Why we Respect our Teachers

A Note on Language Learnability and Active Learning

Purushottam Kar
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
Kanpur, INDIA
purushot@cse.iitk.ac.in

ABSTRACT

Language acquisition - especially in human infants - is a problem that intrigues the layman and baffles the expert. This article takes a computational viewpoint toward the problem and investigates the problem of learnability of languages. We look at some results that show that learning is impossible even under very weak criterion. We next look at a framework in which learning takes place with a helpful teacher and demonstrate the role such a teacher can play in easing the learning problem. The discussion would define language learning formally, survey classical results and point toward recent advances in the field.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Language Acquisition; A.1 [**General Literature**]: INTRODUCTORY AND SURVEY

General Terms

Learning Theory

Keywords

Active Learning, Language Learnability

1. INTRODUCTION

Many of us have witnessed a younger sibling (or a niece/nephew) learn to speak and wondered about it. The wondersome feat achieved by these infants, namely learning a medium of communication used by adults far more experienced and developed neurologically, has captured the attention of researchers for quite some time now. Even the lay person finds himself devoting a minute or two to this problem. If one ponders on the conditions in which language learning takes place then one easily notices a paradox - the infant is simply exposed to utterances in its mother tongue, utterances that are often ill-formed and spontaneous. Most of these utterances are not even directed at the child (in fact all that the child gets directed at itself are mollycoddles - nonsensical blabbers that caretakers make in order to show their affection toward the infant - which, in the author's view, can only obfuscate things further).

Despite being immersed in such a hostile environment, the infant ends up learning the language of its caretakers and eventually becomes a proficient speaker. In order to understand this problem better let us try to pose it in formal terms. Of course this will involve making certain simplifications which we shall reason for as we go along. We shall strive to keep the exposition simple and shall supplement arguments with schematic diagrams to better convey the key ideas.

2. LANGUAGES AND LEARNING

The first simplification that we will make is that of interpreting a language as a set. To see why let us take the set of all words present in our favorite English dictionary which would be sufficient to communicate most intentions - of course our favorite IITK *lingo* would be missing but let us choose to live with this handicap - and call this set Σ. Let Σ^n denote the set of all sentences[1] of length n where the words are taken from Σ. Let $\Sigma^* = \bigcup_{i>0} \Sigma^i$ be the set of all sentences of finite length that use words from Σ. Clearly Σ^* contains all English sentences. However, it also contains sentences like *"This celebrating Jubilee is institute our Golden its year."* and *"How it why now where is."* which are not well formed English sentences. Thus we see that the English "language" can be thought of as that subset $E \subset \Sigma^*$ which contains only well formed grammatical English sentences.[2] For the rest of the article, whenever we refer to a language, it will always be a set $L \subseteq \Sigma^*$.

A language L is said to be **finite** if $|L| < \infty$.[3] Clearly English is not a finite language since given a sentence $s \in E$, one can always create a grammatically correct sentence like *"My friend thinks that s."* of length greater than that of s. Hence I can construct sentences of arbitrarily large lengths which makes $|E| = \infty$.

So suppose the caretakers speak a language $L_t \subset \Sigma^*$

[1]We shall use the terms "string", "utterance" and "sentence" interchangeably in this discussion.

[2]Of course the debate on whether to consider sentences like *"Colorless green ideas sleep furiously."* which although grammatically correct, do not make any sense (or do they?) can be waged here and the author invites readers to wage these debates among themselves.

[3]For a set S, $|S|$ denotes its cardinality: loosely speaking, the number of elements in S.

which the infant must identify or approximate in some sense. What the infant receives is a *finite* number of utterances $s_1, s_2, \ldots, s_n$ where each $s_i \in L_t$. We will call such a sequence a **finite text**. This is a reasonable assumption since the infant seldom receives ill-formed utterances which are tagged as ill-formed [Niy06]. Now the job of the infant is to, given a finite text, identify the **target language** L_t. Suppose the infant has no prior information about the nature or properties of L_t. All it knows is that L_t is some language in Σ^* which contains all the utterances that it has just heard. In this case the infant is faced with the following dilemma - there are an infinite number of such languages: which one should the infant identify as the target?[4] We shall return to this question in a short while (in Section 2.2) after building some more notational apparatus to better discuss the problem.

2.1 Languages and their Grammars

A missing detail in the above discussion involves representation. Since we have already agreed that English (or for that matter any language that supports recursive embeddings - in particular all natural, i.e. human, languages) is an infinite set, representation becomes a problem. In other words how does the infant represent the infinite set it has learnt - it certainly cannot store the entire set explicitly. However we have an intuitive solution to this. All of us have a representation of English in our minds, and a finite one since our minds are finite objects. There do exist several ways of finitely representing infinite sets, two commonly used ones being **automata** and **grammars**. An automaton is like a computer algorithm which can accept input and give output. An automata corresponding to a language is simply an algorithm that answers YES if and only if it is given a string in that language. A grammar, on the other hand, is a set of rules that can be used to generate some strings. A grammar corresponding to a language generates all and only strings in that language.[5]

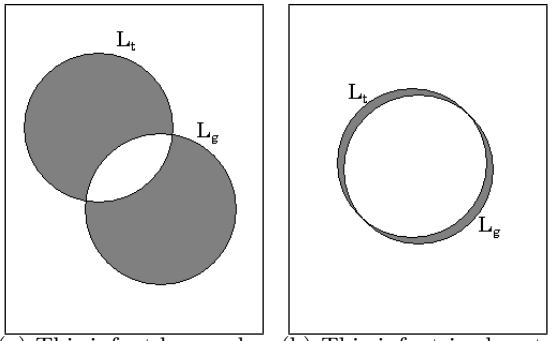
For example take the following language over binary strings $L_1 = \{0^n 1110^m \mid n, m \geq 0\}$. It is a **Regular Language** generated by the following grammar which is a **Regular Expression** $G_1 = 0^*1110^*$. This grammar generates strings which consist of some (or possibly no) zeros followed by three ones followed by some (or possibly no) zeros. It is clear that G_1 generates L_1. It is a simple task to write an algorithm that answers YES if and only if given a string in L_1.

Take the following **Context-free Language** $L_2 = \{0^n 1^n \mid n > 0\}$. This is generated by the following **Context-free Grammar** G_2

$$\begin{aligned} S &\rightarrow 0S1 \\ S &\rightarrow 01 \end{aligned}$$

This grammar generates the string 01 and for every string s that can be generated, the grammar also generates $0s1$. Thus 01 can be generated which in turn paves way for the generation of 0011, and so on. Again it is clear that G_2 generates L_2. It is easy to write an algorithm to say YES on strings in L_2 and No to others.

Whether humans use grammars, algorithmic procedures or some other means to represent languages in their minds is a matter of intense study in a very exciting field called Cognitive Linguistics. However for us it is sufficient that the infant have ways to posit a hypothesis (i.e. its guess of what L_t is) effectively. Given a grammar g hypothesized by the infant, let L_g be the corresponding language. For purposes of evaluation let us assume that we have a notion of *distance* between languages. Thus given two languages L_1 and L_2, we have a distance measure $d : (L_1, L_2) \longmapsto \mathbb{R}$. Many such distance measures can be considered, a natural one being a measure that depends on the symmetric difference of the two languages when interpreted as sets i.e. $L_1 \triangle L_2 = L_1 \backslash L_2 \cup L_2 \backslash L_1$. This distance measure would penalize the infant if it learns a grammar that classifies a large portion of L_t as ungrammatical and a large chunk outside of L_t as well-formed (see Figure 1). One can be even stricter and define $d(L_1, L_2) = 0$ if and only if $L_1 = L_2$ and 1 otherwise.



(a) This infant has probably learnt French (b) This infant is close to learning English though

Figure 1: A distance measure between languages, The shaded portion is $L_t \triangle L_g$

A language L_t will be said to have been learnt on a finite text τ (consisting of strings from L_t) as per a distance measure d if the infant (assuming it starts off with an "initial" hypothesis g_0 corresponding to the language L_0) outputs a grammar g_τ on being exposed to τ such that $d(L_{g_\tau}, L_t) = 0$. A language that can be learnt on any given finite text (so long as the text contains strings from L_t alone) is said to be **learnable**. A class of languages $\mathcal{L}$ (a class of languages is simply a set of languages) is said to be learnable if each $L \in \mathcal{L}$ is learnable (see Figure 2).

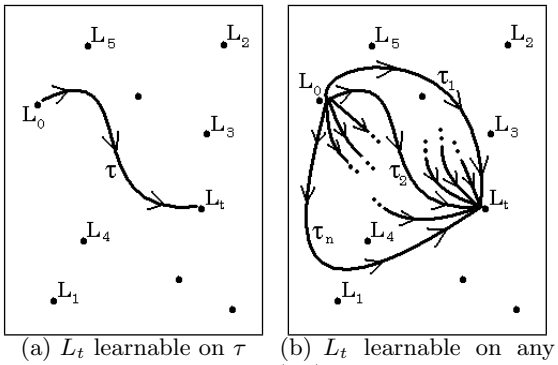
2.2 Language Learnability

Let us formalize the dilemma faced by the infant discussed earlier. The infant is provided with a finite text and has to posit a language as its hypothesis. The problem for the infant is that the target language could be any language that contains the strings it received. In other words, if the infant were to be given an assurance before learning started that the language it has to learn will only come from a **Target Language Class** $\mathcal{L}$, then the class in this case is $\mathcal{L} = 2^{\Sigma^*}$ which in effect gives the infant no apriori information about L_t.[6] One might wonder how such "assurances" can be given

[4]Note that we are assuming that the infant knows Σ. This is a simplifying assumption but is not too unreasonable as concept and word learning predate syntax acquisition[Pin90] although these are not distinct stages.

[5]Under the Church-Turing Hypothesis, only *recursively enumerable sets* admit such finite representations - but we do not have to worry about this technicality - all our languages will be far from raising recursive enumerability questions.

[6]Recall that for any set X, 2^X denotes the *power set* con-



(a) L_t learnable on τ (b) L_t learnable on any text

Figure 2: Learnability of Languages

to an infant. It turns out that if one believes in the *Universal Grammar Hypothesis* [Cho65], then such an assurance is inbuilt in all of us. The hypothesis, very broadly speaking, states that certain universal properties are shared by grammars of all human languages.

Coming back to our problem, the set of languages in 2^{Σ^*} which contain the finite text received by the infant (this holds for any finite text) is vast and these languages are very different from each other according to the distance measures discussed earlier (in fact they would differ widely as per any distance measure that encodes the *generalization* performance of the infant's learnt grammar). Hence the infant has no surety of arriving at a grammar that even closely approximates the target language even if it chooses a language that contains all the utterances it has heard. Thus we arrive at the following result:

THEOREM 1. *The language class $\mathcal{L} = 2^{\Sigma^*}$ is not learnable with finite texts.*

2.3 Learning with Infinite Resources

Were we expecting too much from the infant in the earlier section? Can we relax the learning conditions a bit and see if learning can take place? In particular can we give the infant more sentences to learn the language? Can we restrict the target language class so as to increase the chances of arriving at the target language? We shall see in the following discussion that even if we present the entire language to the infant (by giving it an infinite number of sentences) and restrict the target class to one step beyond the trivial, learnability continues to elude us.

First of all let us give the infant **infinite texts**. An infinite text τ for a language L is an infinite sequence of strings $s_1, s_2, \ldots, s_n, \ldots$ all of which are in L such that every element of L appears at least once in L. By τ_k we shall denote the finite text comprising the first k elements of τ.

Let g_{τ_k} be the infant's hypothesis after receiving τ_k for $k > 0$. Then we say that a language L_t is learnt on an infinite text τ as per a distance measure d *in the limit* if $\lim_{k\to\infty} d(L_{g_{\tau_k}}, L_t) = 0$ i.e. if the infant converges to the target in the limit. Similarly we define what it means for a language and a class of languages to be identifiable in the limit (see Figure 3).

taining all subsets of X, including the empty one. Thus the set 2^{Σ^*} contains all languages. Actually we are just concerned with languages that admit finite representations but beg to gloss over this point.

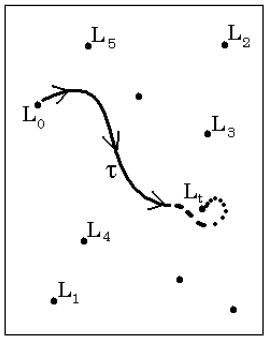


Figure 3: The infant will eventually converge to L_t

Clearly these conditions are weaker than those in Theorem1. However in a seminal paper, Gold [Gol67] demonstrated that even under these weakened conditions and increased resources not only the does language class $\mathcal{L} = 2^{\Sigma^*}$ continue to be non-learnable but the non-learnability persists even the infant is given some prior knowledge about the target language by restricting the target class. We do not give Gold's original proof here but one that follows from results by Blum and Blum [BB75] in a manner presented in [Niy06].

THEOREM 2 (LOCKING TEXT THEOREM). *A language L_t is learnable only if for every $\epsilon > 0$ there exists a finite "locking" text τ_ϵ composed of strings in L_t such that $d(L_{g_{\tau_\epsilon}}, L_t) < \epsilon$ and for all finite texts σ composed of strings in L_t, $d(L_{g_{(\tau_\epsilon \circ \sigma)}}, L_t) < \epsilon$ where $\tau_\epsilon \circ \sigma$ is the concatenation of the two finite sequences.*

Essentially, the theorem says that in order for a language to be learnable, there must exist finite texts that take the infant ϵ close to the target and "lock" it there. Thus, after viewing the locking text, no matter what subsequent utterances it observes, the infant never makes a subsequent hypothesis that is farther off than ϵ i.e. no further exposure can mislead it. We shall prove the theorem by contradiction. We shall show that if locking texts do not exist then we can construct an infinite text on which the infant will never converge to the target. Since convergence is necessary on every infinite text for a language to be called learnable, we shall have proved the theorem.

PROOF. (Sketch) Now for the actual proof. Notice that if there did not exist locking texts for every $\epsilon > 0$, it means for some $\epsilon^* > 0$ there is no corresponding locking text, that is to say no finite text τ taking the infant ϵ^*-close to the target is able to lock it there. Thus for every such text τ that takes the infant close to the target, there must exist a "violator" text σ such that although $d(L_{g_\tau}, L_t) < \delta$, after encountering σ, $d(L_{g_{(\tau \circ \sigma)}}, L_t) > \delta$. See Figure 4 for a schematic.[7]

We can use these violator texts to create an infinite text ζ for which $\lim_{k\to\infty} d(L_{g_{\zeta_k}}, L_t) \neq 0$. We do the following: whenever we observe the infant getting δ-close to the target on a finite text, we feed the infant the violator text corresponding

[7]Note that if there is no locking text for δ then there are none for any $\epsilon < \delta$ either.

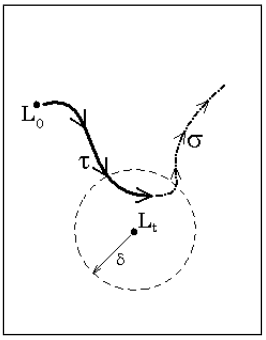


Figure 4: τ brings the infant close but the violator σ spoils the show - i.e. τ cannot be a locking text for δ

to the text the infant has seen until now to force it to give a hypothesis that is at least δ far off from L_t.

For example if the infant gets δ-close on a finite text τ_1, feed it the corresponding violator text (say σ_1) so that $d(L_{g_{(\tau_1 \circ \sigma_1)}}, L_t) > \delta$. Now it is possible that after listening to some more utterances (in the form of another finite text τ'), the infant again comes close to L_t. Let us call the text seen until now τ_2 i.e. let $\tau_2 = \tau_1 \circ \sigma_1 \circ \tau'$. This means $d(L_{g_{\tau_2}}, L_t) < \delta$. But there is no reason to worry because even for τ_2 there would exist some violator text σ_2 (since **no** text can lock the infant to a close neighborhood of L_t) which we will next feed the infant and again take it far away from L_t. Hence we have $d(L_{g_{(\tau_2 \circ \sigma_2)}}, L_t) > \delta$. See Figure 5 for a schematic.

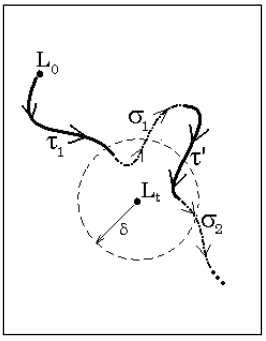


Figure 5: Each time the infant tries to perform well, we can make it perform badly since the infant is not able to lock its good performance

Thus at each step, we are assured of the existence of violator texts since there are no locking texts. This way the infant would at best constantly oscillate in and out of the δ-neighborhood of L_t and can never converge to L_t. □

Thus in order for a language to be learnable (given any text), there must exist finite texts that take the infant arbitrarily close to the target and lock it there. However notice that the existence of such locking texts does not guarantee learnability - it is just that their absence negates any possibility of learning.

This immediately gives us Gold's celebrated result.

THEOREM 3. *Any language class $\mathcal{L}$ that contains all finite languages and at least one infinite language is not learnable in the limit with infinite texts.*

PROOF. (Sketch) Consider such a family $\mathcal{L}$ and an infinite language $L_\infty \in \mathcal{L}$. Since $\mathcal{L}$ is learnable, there must exist finite locking texts for L_∞ for every $\epsilon > 0$. In particular consider the one corresponding to $\epsilon = \frac{1}{2}$ and call it τ. Note that the locking sets themselves are finite languages and hence are contained in $\mathcal{L}$ (since $\mathcal{L}$ contains all finite languages). Thus $\mathcal{L}$ contains L_τ, the set of strings in τ.

Now suppose the infant wants to learn L_τ and the infinite text it gets starts with τ itself. Then we have a problem: although the infant wanted to learn L_τ, it will get locked to L_∞ as τ is a locking text for L_∞. Thus the infant cannot learn L_τ in the limit and hence $\mathcal{L}$ is not learnable in the limit as it contains a language that is not learnable in the limit. □

A little analysis will tell us that in the above situation, although we ended up proving that a finite language is not learnable, it is actually the infinite language that is the trouble maker since finite languages are trivial to learn using just finite texts. However if infinite languages are a problem then we are in a fix since our English language is an infinite one and is believed to be a part of a language class called Context-Free Languages[8] which unfortunately contains all finite languages and also contains English, an infinite language.

The same holds true for the class of Regular Languages which is arguably the simplest possible non-trivial (read interesting) class of languages. Hence we have the following result that dashes all hopes of learnability for interesting language classes.

THEOREM 4. *The classes of Regular and Context-Free languages are not learnable in the limit.*

2.4 Approximately Learning Languages

For those who consider this to be as bad as things can get, the author apologizes for providing yet another set of relaxations which fail to make these language classes learnable. Even if one does not expect the infant to learn the target language exactly but learn some nice approximation (i.e. a grammar that will be correct say 90% of the time), then even learning such an approxmation in a reasonable amount of time is impossible if one believes that a certain mathematical conjecture holds [KV94]. However this is a much more difficult result to prove and we do not attempt to even state the result formally, let alone prove it.

However the mathematical conjecture in itself is fairly interesting. The conjecture is called the Discrete Cube Roots Assumption and it essentially says that a certain invertible function is easy to compute but hard to invert.[9] What makes this conjecture interesting is that all secure Internet communication places faith on the its validity. If this conjecture proves to be false then none of the secure protocols in use today would be secure anymore and attackers would easily be

[8]There is some degree of context-sensitivity in English but we yet again choose not to address this issue further.

[9]Think of the multiplication function that takes two primes and outputs their product. Multiplication is easy but factorization is not.

able to decode encrypted data that is sent over the Internet easily.

Thus the joy of being able to learn in an approximate sense would have to be accompanied by a realization that the next time we key in our credit card details into a payment portal, it would be very simple for an attacker to get hold of all the details.

3. ENTER THE TEACHER

It turns out all that we need to get rid of the non-learnability results given in the previous section is the presence of a teacher! A teacher who can provide answers to certain special types of queries made by the learner, facilitates learning to the extent that it can take place not only in a finite number of steps but actually in a fairly small number of steps.

These results were presented in the seminal papers of [Ang87, Sak90] who proved respectively that the classes of Regular and Context-free languages are learnable with the help of teachers capable of answering two types of queries :

1. Membership Queries : The learner gives the teacher a string s and asks whether $s \in L_t$ or not. The teacher replies back with a YES/NO.

2. Equivalence Queries : The learner gives the teacher its current hypothesis grammar g and asks whether $L_g = L_t$ or not. The teacher either answers YES (in which case the learner is done) or gives a *counterexample* string $s \in L_g \triangle L_t$.

These results outline learning algorithms which when presented with the problem of learning a target language L_t, start asking questions to the teacher. The algorithms process the replies given by the teacher and formulate hypotheses and new questions to ask. If on an equivalence query, the teacher replies back with a YES, then the algorithms halt. It turns out that if the grammar target language L_t is encoded by the grammar g_t, then there exists a universal constant $c > 0$ such that these algorithms do not take more than $|g_t|^c$ steps to converge to the target grammar.[10]

These results, although stimulating, are fairly involved and well beyond the scope of this article. However we do realize the importance of teachers in learning situations such as these (and also in our real lives). Since these results came up, researchers have improved upon them and made them more amenable to practical application. For example we now have genetic algorithms [SPMF01], greedy algorithms [ZL78, Wel84] and kernel based algorithms [CKM07][11] for grammatical inference.

There has also been a lot of research on child language development and although very far from having the final word on child language acquisition, we now have a better idea of how human infants form word-concept correlations and acquire syntax. However this topic merits a dedicated article and we conclude this one with a vote of thanks to all our teachers for making the learning process fun and simple.

[10] $|g|$ denotes the size of the grammar g - i.e. how much space does it take to write down the rules of the grammar.

[11] See Purushottam Kar. An Introduction to Support Vector Machines and their Applications. *Notes on Engineering Research and Development*, xx(yy):pp–qq, 2009. for a discussion on kernel based algorithms.

4. ACKNOWLEDGMENTS

The author thanks Professor Achla M. Raina, Charles Benjamin Strauber and Vedula Vijaya Saradhi for comments on an earlier version of the paper

5. REFERENCES

[Ang87] Dana Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75:87–106, 1987.

[BB75] Manuel Blum and Lenore Blum. Towards a Mathematical Theory of Inductive Inference. *Information and Control*, 28:125–155, 1975.

[Cho65] Noam Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, 1965.

[CKM07] Corinna Cortes, Leonid Kontorovich, and Mehryar Mohri. Learning languages with rational kernels. In *20^{th} Annual Conference on Learning Theory*, 2007.

[Gol67] E. Mark Gold. Language Identification in the Limit. *Information and Control*, 10(5):447–474, 1967.

[KV94] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

[Niy06] Partha Niyogi. *The Computational Nature of Language Learning and Evolution*. The MIT Press, 2006.

[Pin90] Steven Pinker. *Language Acquisition. In an Invitation to Cognitive Science: Language*, volume 1. The MIT Press, 1990.

[Sak90] Yasubumi Sakakibara. Learning Context-free Grammars from Structural Data in Polynomial time. *Theoretical Computer Science*, 76:223–242, 1990.

[SPMF01] Ingo Schröder, Horia F. Pop, Wolfgang Menzel, and Kilian A. Foth. Learning Grammar Weights Using Genetic Algorithms. In *Recent Advances in Natural Language Processing*, 2001.

[Wel84] T. A. Welch. A Technique for High-Performance Data Compression. *Computer*, 17(6):8–19, 1984.

[ZL78] Jacob Ziv and Abraham Lempel. Compression of Individual Sequences Via Variable-Rate Coding. *IEEE Transactions on Information Theory*, 1978.