Another look at indirect negative evidence

Alexander Clark
Department of Computer Science
Royal Holloway, University of London
alexc@cs.rhul.ac.uk

Shalom Lappin
Department of Philosophy
King's College, London
shalom.lappin@kcl.ac.uk

Abstract

Indirect negative evidence is clearly an important way for learners to constrain overgeneralisation, and yet a good learning theoretic analysis has yet to be provided for this, whether in a PAC or a probabilistic identification in the limit framework. In this paper we suggest a theoretical analysis of indirect negative evidence that allows the presence of ungrammatical strings in the input and also accounts for the relationship between grammaticality/acceptability and probability. Given independently justified assumptions about lower bounds on the probabilities of grammatical strings, we establish that a limited number of membership queries of some strings can be probabilistically simulated.

1 Introduction

First language acquisition has been studied for a long time from a theoretical point of view, (Gold, 1967; Niyogi and Berwick, 2000), but a consensus has not emerged as to the most appropriate model for learnability. The two main competing candidates, Gold-style identification in the limit and PAC-learning both have significant flaws.

For most NLP researchers, these issues are simply not problems: for all empirical purposes, one is interested in modelling the distribution of examples or the conditional distribution of labels given examples and the obvious solution – an $\epsilon - \delta$ bound on some suitable loss function such as the Kullback-Leibler Divergence – is sufficient (Horning, 1969; Angluin, 1988a). There may be some complexity issues involved with computing these approximations, but there is no debate about the appropriateness of the learning paradigm.

However, such an approach is unappealing to linguists for a number of reasons: it fails to draw a distinction between grammatical and ungrammatical sentences, and for many linguists the key data are not the "performance" data but rather the "voice of competence" as expressed in grammaticality and acceptability judgments. Many of the most interesting sentences for syntacticians are comparatively rare and unusual and may occur with negligible frequency in the data.

We do not want to get into this debate here: in this paper, we will assume that there is a categorical distinction between grammatical and ungrammatical sentences. See (Schütze, 1996) for extensive discussion.

Within this view learnability is technically quite difficult to formalise in a realistic way. Children clearly are provided with examples of the language – so-called positive data – but the status of examples not in the language – negative data – is one of the endless and rather circular debates in the language acquisition literature (Marcus, 1993). Here we do not look at the role of corrections and other forms of negative data but we focus on what has been called indirect negative evidence (INE). INE is the non-occurrence of data in the primary linguistic data; informally, if the child does not hear certain ungrammatical sentences, then by their absence the child can infer that those strings are ungrammatical.

Indirect negative evidence has long been recognised as an important source of information (Pinker, 1979). However it has been surprisingly difficult to find an explicit learning theoretic account of INE. Indeed, in both the PAC and IIL paradigms it can be shown, that under the standard assumptions, INE cannot help the learner. Thus in many of these models, there is a sharp and implausible distinction between learning paradigms where the learner is provided systematically with every negative example, and those where the learner is denied any negative evidence at all. Neither of these is very realistic.

In this paper, we suggest a resolution for this conflict, by re-examining the standard learnability assumptions. We make three uncontroversial ob-

Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition, pages 26–33,
Athens, Greece, 31 March 2009.

servations: first that the examples the child is provided with are *unlabelled*, secondly that there are a small proportion of ungrammatical sentences in the input to the child, and thirdly that in spite of this, the child does in fact learn.

We then draw a careful distinction between probability and grammaticality and propose a restriction on the class of distributions allowed to take account of the fact that children *are* exposed to some ungrammatical utterances. We call this the Disjoint Distribution Assumption: the assumption that the classes of distributions for different languages must be disjoint. Based on this assumption, we argue that the learner can infer lower bounds on the probabilities of grammatical strings, and that using these lower bounds allow a probabilistic approximation to membership queries of some strings.

On this basis we conclude that the learner does have some *limited* access to indirect negative evidence, and we discuss some of the limitations on this data and the implications for learnability.

2 Background

The most linguistically influential learnability paradigm is undoubtedly that of Gold (Gold, 1967). In this paradigm the learner is required to converge to exactly the right answer after a finite time. In one variant of the paradigm the learner is provided with only positive examples, and must learn on every presentation of the language. Under this paradigm no suprafinite class of languages is learnable. If alternatively the learner is provided with a presentation of labelled examples, then pretty much anything is learnable, but clearly this paradigm has little relevance to the course of language acquisition.

The major problem with the Gold positive data paradigm is that the learner is required to learn under every presentation; given the minimal constraints on what counts as a presentation, this results in a model which is unrealistically hard. In particular, it is difficult for the learner to recover from an overly general hypothesis; since it is has only positive examples, such a hypothesis will never be directly contradicted.

Indirect negative evidence is the claim that the absence of sentences in the PLD can allow a learner to infer that those sentences are ungrammatical. As (Chomsky, 1981, p. 9) says:

> 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.

While this informal argument has been widely accepted, and is often appealed to, it has so far not been incorporated explicitly into a formal model of learnability. Thus there are no learning models that we are aware of where positive learning results have been achieved using indirect negative evidence. Instead positive learnability results have typically used general probabilistic models of convergence without explicitly modelling grammaticality.

In what follows we will use the following notation. Σ is a finite alphabet, and Σ^* is the set of all finite strings over Σ. A (formal) language L is a subset of Σ^*. A distribution D over Σ^* is a function p_D from Σ^* to $[0, 1]$ such that $\sum_{w \in \Sigma^*} p_D(w) = 1$. We will write $\mathcal{D}(\Sigma^*)$ for the set of all distributions over Σ^*. The support of a distribution D is the set of strings with positive probability $supp(D) = \{w | p_D(w) > 0\}$.

3 Probabilistic learning

The solution is to recognise the probabilistic nature of how the samples are generated. We can assume they are generated by some stochastic process. On its own this says nothing – anything can be modelled by a stochastic process. To get learnability we will need to add some constraints.

Suppose the child has seen thousands of times sentences of the type "I am AP", and "He is AP" where AP is an adjective phrase, but he has never heard anybody say "He am AP". Intuitively it seems reasonable in this case to assume that the child can infer from this that sentences of the form"He am AP" are ungrammatical. Now, in the case of the Gold paradigm, the child can make no such inference. No matter how many millions or trillions of times he has heard other examples, the Gold paradigm does not allow any inference to be made from frequency. The teacher, or environment, is an adversary who might be deliberately withholding this data in order to confuse the

learner. The learner has to ignore this information.

However, in a more plausible learning environment, the learner can reason as follows. First, the number of times that the learner has observed sentences of the form "He am AP" is zero. From this, the learner can infer that sentences of this type are rare: i.e. that they are not very probable. Similarly from the high frequency of examples of the type "I am AP" and so on in the observed data, the learner can infer that the probability of these sentences is high.

The second step is that the learner can conclude from the difference in probability of these two similar sets of sentences, that there must be a difference in grammaticality between "He am AP" and "He is AP", and thus that sentences of the type "He am AP" are ungrammatical.

It is important to recognise that the inference proceeds in two steps:

1. the first is the inference from low frequency in the observed data to low probability and
2. the second is the inference from *comparatively* low probability to ungrammaticality.

Both of these steps need justification, but if they are valid, then the learner can extract evidence about what is *not* in the language from stochastic evidence about what *is* in the language. The first step will be justified by some obvious and reasonable probabilistic assumptions about the presentation of the data; the second step is more subtle and requires some assumptions about the way the distribution of examples relates to the language being learned.

3.1 Stochastic assumptions

The basic assumption we make is that the samples are being generated randomly in some way; here we will make the standard assumption that each sentence is generated independently from the same fixed distribution, the Independently and Identically Distributed (IID) assumption. While this is a very standard assumption in statistics and probability, it has been criticised as a modelling assumption for language acquisition (Chater and Vitányi, 2007).

Here we are interested in the acquisition of syntax. We are therefore modelling the dependencies between words and phrases in sentences, but assuming that there are *no* dependencies between different sentences in discourse. That is to say, we assume that the probability that a child hears a particular sentence does not depend on the previously occurring sentence. Clearly, there are dependencies between sentences. After questions, come answers; a polar interrogative is likely to be followed by a "yes" or a "no"; topics relate consecutive sentences semantically, and numerous other factors cause inter-sentential relationships and regularities of various types. Moreover, acceptability does depend a great deal on the immediate context. "Where did who go?" is marginal in most contexts; following "Where did he go?" it is perfectly acceptable. Additionally, since there are multiple people generating Child Directed Speech (CDS), this also introduces dependencies: each person speaks in a slightly different way; while a relative is visiting, there will be a higher probability of certain utterances, and so on. These correspond to a violation of the "identically" part of the IID assumption: the distribution will change in time.

The question is whether it is legitimate to neglect these issues in order to get some mathematical insight: do these idealising assumptions critically affect learnability? All of the computational work that we are aware of makes these assumptions, whether in a nativist paradigm, (Niyogi and Berwick, 2000; Sakas and Fodor, 2001; Yang, 2002) or an empiricist one (Clark and Thollard, 2004). We do need to make *some* assumptions, otherwise even learning the class of observed natural languages would be too hard. The minimal assumptions if we wish to allow any learnability under stochastic presentation are that the process generating the data is stationary and mixing. All we need is for the law of large numbers to hold, and for there to be rapid convergence of the observed frequency to the expectation. We can get this easily with the IID assumption, or with a bit more work using ergodic theory. Thus in what follows we will make the IID assumption; effectively using it as a place-holder for some more realistic assumption, based on ergodic processes. See for example (Gamarnik, 2003) for a an extension of PAC analysis in this direction. The inference from low frequency to low probability follows from the minimal assumptions, specifically the IID, which we are making here.

4 Probability and Grammaticality

We now look at the second step in the probabilistic inference: how can the child go from low probabil-

ity to ungrammaticality? More generally the question is what is the relation between probability and grammaticality. There are lots of factors that affect probability other than grammaticality: length of utterance, lexical frequency, semantic factors and real world factors all can have an impact on probability.

Low probability on its own cannot imply ungrammaticality: if there are infinitely many grammatical sentences then there cannot be a lower bound on the probability: if all grammatical sentences have probability at least ϵ then there could be at most $1/\epsilon$ grammatical sentences which would make the language finite. A very long grammatical sentence can have very low probability, lower than a short ungrammatical sentence, so a less naive approach is necessary: the key point is that the probability must be *comparatively* low.

Since we are learning from unlabelled data, the only information that the child has comes from from the distribution of examples, and so the distribution must pick out the language precisely. To see this more clearly, suppose that the learner had access to an "Oracle" that would tell it the true probability of any string, and has no limit on how many strings it sees. A learner in this unrealistic model is clearly more powerful than any learner that just looks at a finite sample of the data. If this learner could not learn, then no real learner could learn on the basis of finite data.

More precisely for any language L we will have a corresponding set of distributions $\mathcal{D}(L)$, and we require the learner to learn under any of these distributions. What we require is that if we have two distinct languages L and L' then the two sets of distributions $\mathcal{D}(L)$ and $\mathcal{D}(L')$ must be disjoint, i.e. have no elements in common. If they did have a distribution D in common, then no learner could tell the two languages apart as the information being provided would be identical. Of course, given two distinct languages L and L', it is possible that they intersect, that is to say that there are strings w in $L \cap L'$; a natural language example would be two related dialects of the same language such as some dialect of British English and some dialect of American; though the languages are distinct in formal terms, they are not disjoint, as there are sentences that are grammatical in both. When we consider the sets of distributions that are allowed for each language $\mathcal{D}(L)$ and $\mathcal{D}(L')$, we may find that there are elements $D \in \mathcal{D}(L)$ and $D' \in \mathcal{D}(L')$, whose supports overlap, or even whose supports are identical, $supp(D) = supp(D')$, and we may well find that there are even some strings whose probabilities are identical; i.e. there may be a string w such that $p_D(w) = p_{D'}(w) > 0$. But what we do not allow is that we have a distribution D that is an element of both $\mathcal{D}(L)$ and $\mathcal{D}(L')$. If there were such an element, then when the learner was provided with samples drawn from this distribution, since the samples are unlabelled, there is absolutely no way that the learner could work out whether the target was L or L'; the distribution would not determine the language. Therefore there must be a function from distributions to languages. We cannot have a single distribution that could be from two different languages. Let's call this the disjoint distribution assumption (DDA): the assumption that the sets of distributions for distinct languages are disjoint.

Definition 1 *The Disjoint Distribution Assumption: If $L \neq L'$ then $\mathcal{D}(L) \cap \mathcal{D}(L') = \emptyset$.*

This assumption seems uncontroversial; indeed every proposal for a formal probabilistic model of language acquisition that we are aware of makes this assumption implicitly.

Now consider the convergence criterion: we wish to measure the error with respect to the distribution. There are two error terms, corresponding to false positives and false negatives. Suppose our target language is T and our hypothesis is H. Define $P_D(S)$ for some set S to be $\sum_{w \in S} p_D(s)$.

$$e^+ = P_D(H \setminus T) \tag{1}$$

$$e^- = P_D(T \setminus H) \tag{2}$$

We will require both of these error terms to converge to zero rapidly, and uniformly, as the amount of data the learner has increases.

5 Modelling the DDA

If we accept this assumption, then we will require some constraints on the sets of distributions. There are a number of ways to model this: the most basic way is to assume that strings have probability greater than zero if and only if the string is in the language. Formally, for all D in $\mathcal{D}(L)$

$$p_D(w) > 0 \Leftrightarrow w \in L \tag{3}$$

Here we clearly have a function from distributions to languages: we just take the support of the

distribution to be the language: for all D in $\mathcal{D}(L)$, $supp(D) = L$. Under this assumption alone however, indirect negative evidence will not be available.

That is because, in this situation, low probability does not imply ungrammaticality: only *zero* probability implies ungrammaticality. The fact that we have never seen a sentence in a finite sample of size n means that we can say that it is likely to have probability less than about $1/n$, but we cannot say that its probability is likely to be zero. Thus we can never conclude that a sentence is ungrammatical, if we make the assumption in Equation 3, and assume that there are no other limitations on the set of distributions. Since we have to learn for any distribution, we must learn even when the distribution is being picked adversarially. Suppose we have never seen an occurrence of a string; this could be because the probability has been artificially lowered to some infinitesimal quantity by the adversary to mislead us. Thus we gain nothing. Since there is no non-trivial lower bound on the probability of grammatical strings, effectively there is no difference between the requirement $p_D(w) > 0 \Leftrightarrow w \in L$ and the weaker condition $p_D(w) > 0 \Rightarrow w \in L$.

But this is not the only possibility: indeed, it is not a very good model at all. First, the assumption that ungrammatical strings have zero probability is false. Ungrammatical sentences, that is strings w, such that $w \notin L$, do occur in the environment, albeit with low probability. There are performance errors, poetry and songs, other children with less than adult competence, foreigners and many other potential sources of ungrammatical sentences. The orthodox view is that CDS is "unswervingly well-formed" (Newport et al., 1977): this is a slight exaggeration as a quick look at CHILDES (MacWhinney, 2000) will confirm. However, if we allow probabilities to be non-zero for ungrammatical sentences, and put no other restrictions on the distributions then the learner will fail on everything, since any distribution could be for any language.

Secondly, the convergence criterion becomes vacuous. As the probability of ungrammatical sentences is now zero, this means that $P_D(H \setminus T) = e^+ = 0$, and thus the vacuous learner that always returns the hypothesis Σ^* will have zero error. The normal way of dealing with this (Shvaytser, 1990) is to require the learner to hypothesize a subset of the target. This is extremely undesirable, as it fails to account for the presence of over-generalisation errors in the child – or any form of production of ungrammatical sentences. On the basis of these arguments, we can see that this naive approach is clearly inadequate.

There are a number of other arguments why distribution free approaches are inappropriate here, even though they are desirable in standard applications of statistical estimation (Collins, 2005). First, the distribution of examples causally depends on the people who are uttering the examples who are native speakers of the language the learner is learning and use that knowledge to construct utterances. Second, suppose that we are trying to learn a class of languages that includes some infinite regular language L_r. For concreteness suppose it consists of $\{a^*b^*c^*\}$; any number of a's followed by any number of b's followed by any number of c's. The learner must learn under any distribution: in particular it will have to learn under the distribution where every string except an infinitesimally small amount has the number of 'a's equal to the number of 'b's, or under the distribution where the number of occurrences of all three letters must be equal, or any other arbitrary subset of the target language. The adversary can distort the probabilities so that with probability close to one, at a fixed finite time, the learner will only see strings from this subset. In effect the learner has to learn these arbitrary subsets, which could be of much greater complexity than the language.

Indeed researchers doing computational or mathematical modelling of language acquisition often find it convenient to restrict the distributions in some way. For example (Niyogi and Berwick, 2000), in some computational modelling of a parameter-setting model of language acquisition say

> In the earlier section we assumed that the data was uniformly distributed. ... In particular we can choose a distribution which will make the convergence time as large as we want. Thus the distribution-free convergence time for the three parameter system is infinite.

However, finding an alternative is not easy. There are no completely satisfactory ways of restricting the class of distributions, while maintaining the property that the support of the distribu-

tion is equal to the language. (Clark and Thollard, 2004) argue for limiting the class of distributions to those defined by the probabilistic variants of the standard Chomsky representations. While this is sufficient to achieve some interesting learning results, the class of distributions seems too small, and is primarily motivated by the requirements of the learning algorithm, rather than an analysis of the learning situation.

5.1 Other bounds

Rather than making the simplistic assumption that the support of the distribution must equal the language, we can instead make the more realistic assumption that every sentence, grammatical or ungrammatical, can in principle appear in the input and have non zero probability. In this case then we do not need to require the learner to produce a hypothesis that is a subset of the target, because if the learner overgeneralises, e^+ will be non-zero.

However, we clearly need to add some constraints to enforce the DDA. We can model this as a function from distributions to languages. It is obvious that grammaticality is correlated with probability in the sense that grammatical sentences are, broadly speaking, more likely than ungrammatical sentences; a natural way of articulating this is to say that that there must be a real valued threshold function $g_D(w)$ such that if $p_D(w) > g_D(w)$ then $w \in L$. Using this we define the set of allowable distributions for a language L to be:

$$\mathcal{D}(L, g) = \{D : p_D(w) > g_D(w) \Leftrightarrow w \in L\} \tag{4}$$

Clearly this will satisfy the DDA. On its own this is vacuous – we have just changed notation, but this notation gives us a framework in which to compare some alternatives.

The original assumption that the support is equal to the languages in this framework then just has the simple form $g_D(w) = 0$. The naive constant bound we rejected above would be to have this threshold as a constant that depends neither on D nor on w i.e. for all w , $g_D(w) = \epsilon > 0$. Both of these bounds are clearly false, in the sense that they do not hold for natural distributions: the first because there are ungrammatical sentences with non-zero probability; the second because there are grammatical sentences with arbitrarily low probability. But the bound here need not be a constant, and indeed it can depend both on the distribution D and the word w.

5.2 Functional bound

We now look at variants of these bounds that provide a more accurate picture of the set of distributions that the child is exposed to. Recall that what we are trying to do is to characterise a range of distributions that is large enough to include those that the child will be exposed to. A slightly more nuanced way would be to have this as a very simple function of w, that ignores D, and is just a function of length. For example, we could have a simple uniform exponential model:

$$g_D(w) = \alpha_g \beta_g^{|w|} \tag{5}$$

This is in some sense an application of Harris's idea of equiprobability (Harris, 1991):

> whatever else there is to be said about the form of language, a fundamental task is to state the departures from equiprobability in sound- and word-sequences

Using this model, we do not assume that the learner is provided with information about the threshold g; rather the learner will have certain, presumably domain general mechanisms that cause it to discard anomalies, and pay attention to significant deviations from equiprobability. We can view the threshold g as defining a bound on equiprobability; the role of syntax is to characterise these deviations from the assumption that all sequences are in some sense equally likely.

A more realistic model would depend also on D; for example once could define these thresholds to depend on some simple observable properties of the distribution that could take account of lexical probabilities: more sophisticated versions of this bound could be derived from a unigram model, or a class-based model (Pereira, 2000).

Alternatively we could take account of the prefix and suffix probability of a string: for example, where for some $\alpha < 1$: [1]

$$g_D(w) = \alpha \max_{uv=w} p_D(u\Sigma^*) p_D(\Sigma^* v) \tag{6}$$

6 Using the lower bound

Putting aside the specific proposal for the lower bound g, and going back to the issue of indirect

[1] A prefix is just an initial segment of a string and has no linguistic and similarly for a suffix as the final segment.

negative evidence, we can see that the bound g is the missing piece in the inference: if we observe that a string w has zero frequency in our data set, then we can conclude it has low probability, say p; if p is less than $g(w)$, then the string will be ungrammatical; therefore the inference from low probability to ungrammaticality in this case will be justified.

The bound here is justified independently: given the indubitable fact that there is a non-zero probability of ungrammatical strings in the child's input, and the DDA, which again seems unassailable, together with the fact that learners do learn some languages, it is a logical necessity that there is such a bound. This bound then justifies indirect negative evidence.

It is important to realise how limited this negative evidence is: it does not give the learner unlimited access to negative examples. The learner can only find out about sentences that would be frequent if they were grammatical; this may be enough to constrain overgeneralisation.

The most straightforward way of formalising this indirect negative evidence is with *membership queries* (Valiant, 1984; Angluin, 1988b). Membership queries are a model of learning where the learner, rather than merely passively receiving examples, can query an oracle about whether an example is in the language or not. In the model we propose, the learner can approximate a membership query with high probability by seeing the frequency of an example with a high g in a large sample. If the frequency is low, often zero, in this sample, then with high probability this example will be ungrammatical.

In particular given a functional bound, and some polynomial thresholds on the probability, and using Chernoff bounds we can simulate a polynomial number of membership queries, using large samples of data. Note that membership queries were part of the original PAC model (Valiant, 1984). Thus we can precisely define a limited form of indirect negative evidence.

In particular given a bound g, we can test to see whether a polynomial number of strings are ungrammatical by taking a large sample and examining their frequency.

The exact details here depend on the form of $g_D(w)$; if the bound depends on D in some respect the learner will need to estimate some aspect of D to compute the bound. This corresponds to working out how probable the sentence would be if it were grammatical. In the cases we have considered here, given sufficient data, we can estimate $g_D(w)$ with high probability to an accuracy of ϵ_1; call the estimate $\hat{g}_D(w)$. We can also estimate the actual probability of the string with high probability again with accuracy ϵ_2: let us denote this estimate by $\hat{p}_D(w)$. If $\hat{p}_D(w) + \epsilon_2 < \hat{g}_D(w) - \epsilon_1$, then we can conclude that $p_D(w) < g_D(w)$ and therefore that the sentence is ungrammatical. Conversely, the fact that a string has been observed once does not necessarily mean that it is grammatical. It only means that the probability is non-zero. For the learner to conclude that it is grammatical, s/he needs to have seen it enough times to conclude that the probability is above threshold. This will be if $\hat{p}_D(w) - \epsilon_2 > \hat{g}_D(w) + \epsilon_1$

Note that this may be slightly too weak and we might want to have a separate lower bound for grammaticality and upper bound for ungrammaticality. Otherwise if the distribution is such that many strings are very close to the boundary it will not be possible for the learner to determine whether they are grammatical or not.

We can thus define learnability with respect to a bound g that defines a set of distributions $\mathcal{D}(L, G)$. Thus this model differs from the PAC model in two respects: first the data is unlabelled, and secondly is is not distribution free.

Definition An algorithm A learns the class of languages $\mathcal{L}$ if there is a polynomial p such that for every language $L \in \mathcal{L}$, where n is the size of the smallest representation of L, for all distributions $D \in \mathcal{D}(L, g)$ for all $\epsilon, \delta > 0$, when the algorithm A is provided with at least $p(n, \epsilon^{-1}, \delta^{-1}, \Sigma)$ unlabelled examples drawn IID from D, it produces with probability at least $1 - \delta$ a hypothesis H such that the error $P_D(H \setminus T \cup T \setminus H) < \epsilon$ and furthermore it runs in time polynomial in the total size of the sample.

7 Discussion

The unrealistic assumptions of the Gold paradigm were realised quite early on (Horning, 1969). It is possible to modify the Gold paradigm by incorporating a probabilistic presentation in the data and requiring the learner to learn with probability one. Perhaps surprisingly this does not change anything, if we put no constraints on the target distribution (Angluin, 1988a).

In particular given a presentation on which the

normal non-probabilistic learner fails, we can construct a distribution on which the probabilistic learner will fail. Thus allowing an adversary to pick the distribution is just as bad as allowing an adversary to pick the presentation. However, the distribution free assumption with unlabelled data cannot account for the real variety of distributions of CDS. In this model we propose restrictions on the class of distributions, motivated by the occurrence of ungrammatical sentences. This also means that we do not require a separate bound for over-generalisation. As a result, we conclude that there are limited amounts of negative evidence, and suggest that these can be formalised as a limited number of membership queries, of strings that would occur infrequently if they were ungrammatical.

To be clear, we are not claiming that this is a direct model of how children learn languages: rather we hope to get some insight into the fundamental limitations of learning from unlabelled data by switching to a more nuanced model. Here we have not presented any positive results using this model, but we observe that distribution dependent results for learning regular languages and some context free languages could be naturally modified to learn in this framework. We hope that the recognition of the validity of indirect negative evidence will direct attention away from the supposed problems of controlling overgeneralisation and towards the real problems: the computational complexity of inferring complex models.

References

D. Angluin. 1988a. Identifying languages from stochastic examples. Technical Report YALEU/DCS/RR-614, Yale University, Dept. of Computer Science, New Haven, CT.

D. Angluin. 1988b. Queries and concept learning. *Machine Learning*, 2(4):319–342, April.

N. Chater and P. Vitányi. 2007. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163.

N. Chomsky. 1981. Lectures on Government and Binding.

Alexander Clark and Franck Thollard. 2004. Partially distribution-free learning of regular languages from positive samples. In *Proceedings of COLING*, Geneva, Switzerland.

M. Collins. 2005. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In Harry Bunt, John Carroll, and Giorgio Satta, editors, *New Developments In Parsing Technology*, chapter 2, pages 19–55. Springer.

D Gamarnik. 2003. Extension of the PAC framework to finite and countable Markov chains. *IEEE Transactions on Information Theory*, 49(1):338–345.

E. M. Gold. 1967. Language identification in the limit. *Information and control*, 10(5):447 – 474.

Z.S. Harris. 1991. *A Theory of Language and Information: A Mathematical Approach*. Clarendon Press.

James Jay Horning. 1969. *A study of grammatical inference*. Ph.D. thesis, Computer Science Department, Stanford University.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates Inc, US.

G.F. Marcus. 1993. Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

E.L. Newport, H. Gleitman, and L.R. Gleitman. 1977. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In *Talking to children: Language input and acquisition*, pages 109–149. Cambridge University Press.

Partha Niyogi and Robert C. Berwick. 2000. Formal models for learning in the principle and parameters framework. In Peter Broeder and Jaap Murre, editors, *Models of Language Acquisition*, pages 225–243. Oxford University Press.

F. Pereira. 2000. Formal grammar and information theory: Together again? In *Philosophical Transactions of the Royal Society*, pages 1239-1253. Royal Society, London.

Steven Pinker. 1979. Formal models of language learning. *Cognition*, 7:217–282.

W. Sakas and J.D. Fodor. 2001. The structural triggers learner. In *Language Acquisition and Learnability*, pages 172–233. Cambridge University Press.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics*. University of Chicago Press.

H. Shvaytser. 1990. A necessary condition for learning from positive examples. *Machine Learning*, 5(1):101–113.

L. Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134 – 1142.

C.D. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press, USA.