Formal Models of Language: Formal versus Natural Language

Paula Buttery

Easter 2018

3. Language Learnability

If we define a ***grammatical system***, $(\mathcal{H}, \Omega, \mathcal{L})$ as:

- $\mathcal{H}$ a hypothesis space of language descriptions (e.g. all possible grammars)
- Ω a sample space (e.g. all possible strings)
- $\mathcal{L}$ a function that maps from a member of $\mathcal{H}$ to a subset of Ω

Then a ***learning function***, F, maps from a subset of Ω to a member of $\mathcal{H}$.[1]

[1] For example, if we have $(\mathcal{H}_{cfg}, \Sigma^*, \mathcal{L})$ (that is, the grammatical system of all context-free languages over Σ) then for some $G \in \mathcal{H}_{cfg}$:

- $\mathcal{L}(G) = \{s_a, s_b, s_c ...\} \subseteq \Sigma*$
- and $F(\{s_a, s_b, s_c ...\}) = G$ for some $\{s_a, s_b, s_c ...\} \subseteq \Sigma*$

learnability

Learnability is a property of a language class and occurs when F is surjective (when we can learn every grammar in the hypothesis space using the learning function). The learning function manifests as an algorithm for ***grammar induction*** (and is often referred to as the ***learner***).

Gold's paper on learnability[2] introduced a number of ***learning paradigms*** one of which has been extremely influential in Linguistics, the details are as follows:

[2] E Mark Gold. Language identification in the limit. *Information and Control*, 10 (5):447 – 474, 1967. ISSN 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(67)91165-5. URL `http://www.sciencedirect.com/science/article/pii/S0019995867911655`

For a grammatical system $(\mathcal{G}, \Sigma^*, \mathcal{L})$ –

- An $L \in \mathcal{L}$ is selected as the ***target language*** (i.e. the language that the learner is attempting to learn).
- All samples from L (i.e. all s_i such that $s_i \in L$) are presented to the learner one at a time, $s_1, s_2, ...$, in an infinite sequence.[3]
- After receiving each sample, the learner produces a hypothesis, $G_i \in \mathcal{G}$.[4]
- Learning is *successful* when G has been ***identified in the limit***: that is, there is some number N such that for all $i > N$, the hypothesised grammar $G_i = G_N$, and $\mathcal{L}(G_N) = L$ (the target language).[5]

[3] Note that the learner receives only ***positive evidence*** (as opposed to ***negative evidence*** which would be where strings not in L were also presented to the learner but specifically flagged as errors). Also note that the evidence is exhaustive (i.e. every $s \in L$ will eventually be presented in the sequence.)

[4] So, after seeing the sequence $s_1, ...s_n$, the learner produces G_n.

[5] Note that N is finite but there are no constraints placed on the computation time of the learning function.

In this paradigm the class of languages, G, is learnable if every language in the class can be identified in the limit, no matter what order the samples appear in. A well known result of Gold's work is that suprafinite classes of languages[6] are not learnable.

[6] A suprafinite class of languages is one that contains all possible finite languages and at least one infinite language—all the language classes in the Chomsky hierarchy are suprafinite.

Child language acquisition versus Gold

Gold provides us with a framework for a thought experiment in which specific details must be fleshed out; in particular the defini-

tion of the hypothesis-space, $\mathcal{H}$, and the learning function, F.

Some linguists (sometimes called ***nativists***), believe in innate linguistic knowledge or a specific language faculty in the brain.[7] From the point of view of these linguists the hypothesis-space of grammars is relatively small, constrained by the innate knowledge.[8] Learning functions in this scenario tend to be algorithmic, analysing an input string and moving systematically from one grammar to the next within the small hypothesis-space.

Empirical or ***usage-based*** linguists, on the other hand, believe that language may be acquired without the aid of an innate language faculty. These linguists have suggested that learning can be modelled as a statistical competition between all the grammars within the hypothesis-space. For these linguists, the hypothesis-space is unconstrained and could consequently be very large. A statistical learning function returns a probability distribution over the possible grammars. The distribution represents each grammar's fitness to describe the sentences encountered so far. In this scenario the current hypothesised grammar, G_i, could be selected according to the distribution. Note that under this model of learning, there needs to be a modified definition for success: for example, we could say that F ***converges*** to $G \in \mathcal{H}$ if there exists a finite N such that for all $i > N$, F is defined on $\{s_1 ... s_i\}$ and returns a distribution over $\mathcal{H}$ such that G is most likely.

Notice there are several points of difference between Gold's learning paradigm and language acquisition in children:

1 Gold's paradigm requires convergence in a finite number of steps (i.e after a finite number of hypothesised grammars). The amount of data the learner sees, however, is unbounded and the learner can use unbounded amounts of computation.

- In child language acquisition a child only sees a limited amount of data, and has only limited computational resources.

2 Gold's paradigm doesn't tell us anything about a learner's state at any particular time. In fact, at any particular time, it is not possible to tell whether learning has been successful (identified in the limit), since the learner may always guess a new grammar when presented with the next sentence.

- In reality children learn progressively and could perhaps be considered to be converging towards a target language (as is described above for the statistical learning models).

3 The learner hypothises a grammar after every presentation of a string—this includes presentations that have been chosen by an adversary with knowledge of the internal state of the learner.

- It is arguable that actual input distributions received by children are in some way helpful (referred to as ***parentese***) and that children might even receive helpful negative evidence (as opposed to

[7] This is referred to by Chomsky as ***Universal Grammar***.

[8] Nativists might argue that the hypothesis space *must* be constrained due to Gold's result that *none* of the classes in the Chomsky hierarchy are learnable—whether the learnability of a Chomskyan classes is relevant to a human learner is a matter of debate.

positive evidence only).[9] It has also been suggested that children only attend selectively to evidence—that is, they notice only the strings that are *just right* for them to learn from (this is referred to as the *Goldilocks effect*).

[9] Note that linguists do not agree on these points.

4 Within Gold's paradigm the target language is static and the learner is required to exactly identify the target language.

- Natural languages are dynamic not static. Also some linguists claim that we can observe differences in word choices and grammaticality judgments between adults speakers from quite similar backgrounds (that is, they do not appear to have a common target language). It is also not without argument that we ever converge on a single stable grammar within our lifetimes.

References

E Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447 – 474, 1967. ISSN 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(67)91165-5. URL http://www.sciencedirect.com/science/article/pii/S0019995867911655.