Language Learning Problems in the Principles and Parameters Framework

Partha Niyogi

Presented by Chunchuan Lyu

March 22, 2016

Overview

1. The Principles and Parameters Framework
   - space of learning problems
   - X-bar theory
   - Triggering Learning Algorithm

2. Formal Analysis of the Triggering Learning Algorithm
   - parameters space learning wih TLA as a Markov chain
   - learnability theorem
   - rate of convergence

3. Discussion
   - summary
   - continued works and open questions

The Principles and Parameters Framework: Problem of Language Learning

Question: under what condition is language learning possible?

Goal: put linguistic problem into formal analysis

Given data distribution $\mathcal{P}$, use an algorithm $\mathcal{A}$ to identify target grammar $\mathcal{G}_t$ within Hypothesis space $\mathcal{H}$.

The Principles and Parameters Framework: Space of learning problems

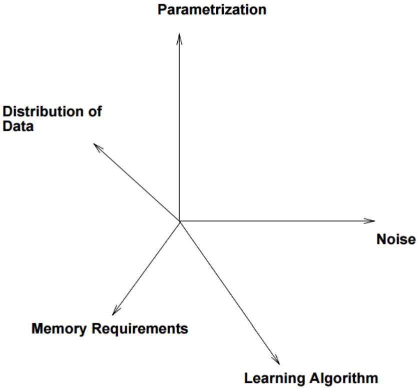

Figure: five important dimensions of every learning problem

The Principles and Parameters Framework: X-bar theory

- Production Rules:

$$XP \rightarrow SpecX'(p_1 = 0) \; or \; X'Spec(p_1 = 1)$$

$$X' \rightarrow CompX'(p_2 = 0) \; or \; X'Comp(p_2 = 1)$$

$$X' \rightarrow X$$

$$Spec/Comp \rightarrow XP \; or \; \epsilon$$

- X: lexical type e.g. Noun,Verb,Adjective
- Spec/Comp : specifier/complement
- p_3 : V2 parameter (binary) control transformation from underlining structure to surface form

3 binary parameters in total, $\mathcal{G}_t$ could be parameterized as [1 0 1]

The Principles and Parameters Framework: X-bar theory

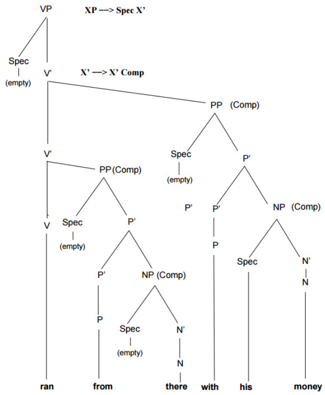

Figure: spec-first, and comp-final

The Principles and Parameters Framework: Triggering Learning Algorithm

1. [Initialize] Start at random initial point in space of possible parameter settings, specifying a single hypothesized grammar $\mathcal{G}_t$ with its' resulting extension as a language $\mathcal{L}(\mathcal{G}_t)$;
2. [Process input sentence] Receive a positive example sentence s_i at time t_i (drawn from distribution over $\mathcal{L}(\mathcal{G}_t)$);
3. [Learnability on error detection] If the current grammar parses s_i, then go to Step 2; otherwise, continue;
4. [Single-step hill climbing] Select a single parameter uniformly at random, to flip from its current setting, and change it iff that change allows the current sentence to be analyzed

Remarks:

- Only change one parameter at one step
- Only change parameter when it parses
- Memoryless

Formal Analysis of the Triggering Learning Algorithm: parameter space learning wih TLA as a Markov chain

- N binary parameters: space $\mathcal{H}$ of grammars with 2^N points
- Transitions between states differ only by one parameter
- p_{ij} defines probability of transition from state i to state j
- Probabilities p_{ij} are determined by a probability distribution $\mathcal{P}$ over $\mathcal{L}(\mathcal{G}_t)$, and the learning algorithm $\mathcal{A}$

$$< \mathcal{A}, \mathcal{P}, \mathcal{G}_t, \mathcal{H} > \text{ characterize a Markov chain.}$$

Formal Analysis of the Triggering Learning Algorithm: parameter space learning wih TLA as a Markov chain

Definition (Gold-learnability)

$\mathcal{G}_t$ is Gold learnable by $\mathcal{A}$ for distribution $\mathcal{P}$ if and only if $\mathcal{A}$ identifies $\mathcal{G}_t$ in the limit of $n \to \infty$ with probability 1.

Definition (Absorbing state)

Given a Markov chain M, and absorbing state of M is a state $s \in M$ that has no exit arcs to any others states of M.

Definition (Closed set of states)

A closed set of states (C) is any proper subset of states in M such that there is no transition probability from any of the states in C to any state not in C.

Formal Analysis of the Triggering Learning Algorithm: learnability theorem

Theorem

Let $< \mathcal{A}, \mathcal{P}, \mathcal{G}_t, \mathcal{H} >$ be a memoryless learning system. Let M be the Markov chain associated with this learning system. Then, $\mathcal{G}_t$ is Gold learnable by $\mathcal{A}$ for distribution $\mathcal{P}$ if and only if every closed set in M includes the target state corresponding to $\mathcal{G}_t$.

Interpretation:

- [only if] Gold-learnability requires algorithm makes no mistake from which it cannot recover.
- [if] as long as it won't make such mistake, Gold-learnability can be assured.

Formal Analysis of the Triggering Learning Algorithm: learnability theorem

Theorem (learnability theorem)

Let $< \mathcal{A}, \mathcal{P}, \mathcal{G}_t, \mathcal{H} >$ be a memoryless learning system. Let M be the Markov chain associated with this learning system. Then, $\mathcal{G}_t$ is Gold learnable by $\mathcal{A}$ for distribution $\mathcal{P}$ if and only if every closed set in M includes the target state corresponding to $\mathcal{G}_t$.

Key ideas in proof (only for expert):

- [only if] Proof by contradiction.
- [if] By construction, M can be decomposed into transient states and the only closed set of state, which only contains target state. As number of sentences turns to infinity, the probability of starting from any transient state to all other transient states turns to 0. This left the limiting probability of identify $\mathcal{G}_f$ to be 1.

Formal Analysis of the Triggering Learning Algorithm: Rate of convergence

How fast is the convergence?
Computing transition probability:

1. [Assign distribution] Fix a probability measure $\mathcal{P}$ on the sentences of target language $\mathcal{L}(\mathcal{G}_t)$ (short hand as $\mathcal{L}_t$) ;
2. [Enumerate states] Assign a state to each language i.e., each $\mathcal{L}_i$;
3. [Normalize by the target language] $\mathcal{L}'_i = \mathcal{L}_i \cap \mathcal{L}_t$;
4. [Take set differences] For any two states i and j, where $i \neq j$, if their corresponding grammar has more than one difference in parameter $p_{ij} = 0$. Otherwise, $p_{ij} = \frac{1}{N} P(\mathcal{L}'_j \backslash \mathcal{L}'_i)$, where N is number of parameters. For $i = j$, we have $p_{ii} = 1 - \sum_{i \neq k} p_{ik}$.

Formal Analysis of the Triggering Learning Algorithm: rate of convergence

Represent transition matrix P with eigenvalue decomposition:

$$P^k = \sum_{i=1}^{m} \lambda_i^k y_i x_i$$

Let Π_0 represents probability over initial states, which could be uniform. Let $\Pi_k = (\pi_1(k), \cdots, \pi_m(k))$ be the probability distribution over states at time k, and $\Pi = \underset{k\to\infty}{\Pi_k} = \Pi_0 P_\infty$.

$\lambda_1 = 1$ correspond to target state, which is an absorbing state.

Theorem (convergence theorem)

$$||\Pi_k - \Pi|| = ||\sum_{i=2}^{m} \lambda_i^k \Pi_0 y_i x_i|| \leq \max_{2 \leq i \leq m} \{|\lambda_i|^k\} \sum_{i=2}^{m} ||\Pi_0 y_i x_i||$$

Interpretation:

- If there are more than one $\lambda_i = 1$, the target grammar is not Gold-learnable. This is to the presence of multiple absorbing states, which implies existence of closed set of states that does not contain the target state.
- Otherwise, the second largest eigenvalue of transition matrix determines the speed of convergence.

Summary

- Turning linguistics problem into formal analysis allows making precise statement
- Learning in parameterized theory – five elements/use the picture

1	parameterization	finite
2	distribution of data	independent
3	noise	noiseless
4	algorithm	TLA
5	memory	memoryless

- Learning in such framework can be framed as a Markov chain
    1. Gold-learnable iff target state is included in every closed set of states
    2. Rate of convergence can be computed from transition matrix

Extended works

- Alternating the five elements. e.g. changing TLA to random walk ensures learnability.
- More formalism, introduce metric between grammar, probably approximately correct, VC and so on.
- Learning over generation can be modeled as population dynamics, thus a framework for language evolution.
- The origin of language, why language work as it is. Treat grammar as genre, and communication efficiency as selection criteria.

Open questions

- How to handle ambiguity in parsing. $\mathcal{L}$ does not seem to be identifiable with $\mathcal{G}$.
- Countable and Uncountable parameterization. The first theorem requires finiteness.
- Interactions between teacher distribution and learner.

Q & A?

References

Partha Niyogi.
The computational nature of language learning and evolution.
MIT press Cambridge, MA:, 2006.

Partha Niyogi.
The informational complexity of learning: perspectives on neural networks and generative grammar.
Springer Science & Business Media, 2012.

Matilde Marcolli.
Cs101: Mathematical and computational linguistics, 2015.