‘Ideal Learning’ of Natural Language: Positive Results about Learning from Positive Evidence

Ray Mendoza
Psych 245A
Thursday, April 24th, 2008

Outline

- Ideal Learner
- Ideal Language Learning by Simplicity
- The Prediction Theorem and Ideal Language Learning
- The Ideal Learning of Grammaticality Judgements
- The Ideal Learning of Language Production
- The Poverty of Stimulus Reconsidered
- Summary

Ideal Learner

- Propose an ideal learner
  - Cannot learn from only positive data
  - Assumes negative linguistic data is non-critical (and non-existent)
- Problem 2: Baker’s Paradox
  - Specific constructions cannot be learned
  - Considers learnability from a construction-specific perspective
- Learner follows Simplicity Principle (i.e. Occam’s Razor)
- Chater/Vitany: “you can learn given an ideal learner, but this does not mean a child can” (child ≠ ideal learner)

Ideal Language Learning by Simplicity

- Four components
  - Class of Linguistic inputs (environment)
  - Class of possible models of language (linguistic structure)
  - Measure of learning performance
  - Learning model
Class of Linguistic Inputs
- Potentially Infinite
- Represented as a binary string
- Produced by a real computational process
- Combined with random input explains effect of non-deterministic input to learner
- Modeled with
  - Monotone Turing machine
  - Random input to the machine
- Random programming monkeys ex.

Class of Possible Models of Language
- Gold’s Theorem: Only provides for identification
- Chomsky’s context-free languages: principles and parameters framework
- The model of linguistic data must be generated by a computable process
  - Rules must have a mechanism by which they are learned
  - Movement must have a mechanism
  - All linguistic constructs must be learned
  - (perhaps this will one day give feedback into linguistic theory!)

Measuring Learning Performance
- Primary measure: prediction
  - Can we predict the continuation of utterances
  - Adds another level of complexity to learning
  \[ \mu_C(0|x) = \mu_C(x0) / \mu_C(x) \] prob. of 0 given x
- Learner doesn’t know true distribution of \( \mu_C \)

The Learning Method: Predicting by Simplicity
- Simplistic explanation of data preferred
- Considers predictions from various hypothesizes
- Applies theory which generates simplest encoding of data (consistent with data)
  - consistency \( \leftrightarrow \) simplicity
- We favor shortest encoding of data (based on Kolmogorov complexity)
The Learning Method: Predicting by Simplicity

- "By using a universal programming language, the learner can be sure to be able, at least in principle, to represent every such computational process."
- Issues
  - Encoding length varies among different 'universal programming' languages
  - Length of encoding has depends on the mental representations (we must presuppose this to begin with!)
- Encodings (choose your poison)
  - phrase structure
  - tree-adj. grammar < Minimalist Program < Govt. & Binding
  - categorical grammar

The Prediction Theorem and Ideal Language Learning

- Prediction Theorem
  - Given a universal monotone distribution
    \( \lambda : \text{universal monotone distribution (target)} \)
    \( \mu : \text{computable monotone distribution (learned)} \)
    \( \lambda(0|x) = \lambda(x)/\lambda(x) \)
    \( \text{prediction} \)
    \( \text{Error (x)} = (\lambda(0|x) - \mu(0|x))^2 \)
    \( S_n = \sum \mu(x) \text{Error (x)} \)
    \( \sum_{j=1}^{\infty} S_j \leq \frac{\log_2 2}{2} K(\mu) \)
- In the limit, error is bounded (and in some sense minimized)

The Ideal Learning of Grammaticality Judgements

- Discusses asymptotic behavior of overgeneralization and undergeneralization
- Learners overgeneralize
  \( \Delta_j(x) = \sum P_i(k|x) \) error prob. of jth symbol
- Does not account for phrasal structure in probabilities
  \( \langle \Delta_j \rangle = \sum P_i(k|x) \Delta_j(x) \) average
  \( \sum \langle \Delta_j \rangle \leq K(\mu) / \log_2 2 \) bounds all avg. err.

The Ideal Learning of Grammaticality Judgements

- Learners undergeneralize in practice
- Under Simplicity Principle, this should not happen
  - More general explanations which account for more data are preferred over special cases which leading to idiosyncrasies
- Soft undergeneralization
  \( \Lambda_j(x) = \sum P_i(k|x) \Lambda_j(x) \) prob. of accurate prediction of prob. dist \( P_i \)
  - Does not account for phrasal structure in probabilities
  \( \langle \Lambda_j \rangle = \sum P_i(k|x) \Lambda_j(x) \) weighted average
  \( \sum \langle \Lambda_j \rangle \leq K(\mu) / \log_2(f/e) \) bound prob. of soft generalization
The Ideal Learning of Language Production

- Acquisition also comprises of production
- Learned distribution approaches univ. dist.

\[ \lambda(y|x) / \mu(y|x) \rightarrow 1 \] conv. of prob. distns

Thus, using learned distribution for production ensures mutual (native) intelligibility

The Poverty of Stimulus Reconsidered

- Constraint-based systems complicate productive grammar, but constraints “can be learned given enough positive data” (needs to be fleshed out)
- Identification in the limit, we (last time) concluded is inapplicable, in general, to language acquisition
  - His problem is identification, not learning
- Statistical properties of language, bivalence of grammaticality judgements (probabilistic models of language reception/production)
- Absence as implicit negative evidence

Summary

- Defined simplistic learner
- Convergence of predictive capabilities
- Convergence of grammaticality judgement
- Convergence of language production
- Language is learnable from positive input

Questions

- Any questions, comments or suggestions?