Research seminar week 4

Tamás Biró

Humanities Computing

University of Groningen

t.s.biro@rug.nl

This week: introduction to learning

Why learning?

Theoretical and psycholinguistic goals:

- Model child language acquisition (L1)
- Model foreign language acquisition (L2)

Language technology:

- Developing software (last approx. 20 years; stating rules not good enough)

Machine learning

- More maths & computing, less linguistics. Most of contemporary computational linguistics. See separate course.

- Typically: categorization task
  - A word: which part-of-speech?
  - An ambiguous word: which meaning?
  - An acoustic sound: which phoneme?
  - A text: what emotional attitude?

Language in a generative approach

- $Alphabet$ = finite set of characters

- $Language$ = (finite/infinite) set of "words" composed of letters in alphabet.

- $Grammar$ = a construction defining which strings belong to the language, and which don't.

Task of learning

- Unknown *target language/grammar*

- Given set of possible languages (*family*) or set of possible grammars.

- Given *learning data*: words in target L.

- Task: using some algorithm, find target language or target grammar.

Gold's theorem

- If family consists of all the finite languages and at least one infinite language, then the family is not learnable in the limit.

Cf. Niyogi, section 2.2; next week.

How do children do?

- Chomsky: innate principles, only need to learn parameter values.
- Kirby and others: learning biases.

Examples for “generalized P&P”

- P&P: given universal principles, find (binary) values of the parameters.
- OT: given universal constraints, find hierarchy (ranking)
- HG: given universal constr, find weights.

NB: often, more grammars describe the same language.

Example for learning biases

1. Given enumeration of languages/grammars: L_1, L_2, etc..

2. Start with $L_i = L_1$.

3. If next learning datum w not in L_i, find next language in the list that contains w. Repeat.

Online and offline learning

- Offline learning algorithm: first collect all data, then find grammar consistent with all/most of them.

- Online learning algorithm: after each data, update your *hypothesis grammar*.

Next week

- Background in general learning algorithms.
- TLA: learning algorithm for P&P.
- RCD, EDCD, GLA: learning algorithms for OT (and HG).

By next week:

- Learning by enumeration in your approach?
- Have a loot at Niyogi chapters 2-4.

Niyogi: expected to be referred to in final paper.
More references to come in coming weeks + ask me!

First student presentation

- First student presentations in 2/3 weeks.
- Each student: cca. 15 minutes with slides.
- 1. Background and model.
  2. State research problems & hypothesis.
  3. Technical details of implementation (why so?).