

LCTG—Notes Part 10: Language Acquisition by Child and Machine

1

Example

- The Stage VI child has encountered a dog. Then she encounters **more dogs**.

(1) a. Child: (thinks:) $more' dog'$

b. Adult: “More doggies!”

c. Child’s lexical candidates:

more:= **NP(X)/N(X) : more'**_{((e,t),e)}

$NP(X)\backslash N(X) : more'$ _{((e,t),e)}

$N(more) : dogs'$ _(e,t)

doggies:= $NP(X)/N(X) : more'$ _{((e,t),e)}

$NP(X)\backslash N(X) : more'$ _{((e,t),e)}

N(doggies) : dogs'_(e,t)

more doggies:= $NP(more\ doggies) : (more' dogs')_e$

- All of these candidates are permitted by the universal lexical principles of UG.

3

Child and Computer Language Development

- The child’s problem is similar to the problem of inducing a treebank grammar, but a little harder.
 - They have **unordered logical forms**, not language-specific ordered derivation trees.
 - So they have to work out **which word(s) go with which element(s) of logical form**, as well as the directionality of the syntactic categories (which are otherwise universally determined by the semantic types of the latter).
- They do not seem to have to deal with a greater amount of error than the Penn WSJ treebank has (McWhinnie 2005).
 - But they may need to deal with **situations which support a number of logical forms**.
 - And they need to be able to recover from temporary **wrong lexical assignments**.
 - And they need to be able to handle **lexical ambiguity**.

2

However, not all of them are consistent with *this* utterance in *this* language. The new utterance “More doggies” supports just three derivations, as follows:

(2) a. $\frac{\text{MORE} \quad \text{DOGGIES} \quad !}{NP(X)/N(X) : more'_{((e,t),e)} \quad N(doggies) : dogs'_{(e,t)}} \rightarrow$

$\frac{\quad \quad \quad}{NP(doggies) : more' dogs'_e}$

b. $\frac{\text{MORE} \quad \text{DOGGIES} \quad !}{N(more) : dogs'_{(e,t)} \quad NP(X)\backslash N(X) : more'_{((e,t),e)}} \leftarrow$

$\frac{\quad \quad \quad}{NP(more) : more' dogs'_e}$

c. $\frac{\text{MORE DOGGIES} \quad !}{NP(more\ doggies) : more' dogs'_e}$

- The other candidates may or may not be supported by later utterances, but the present utterance gives no information on them. They are therefore dropped from further consideration in this cycle, leaving the following reduced set of candidates:

4

(3) The child's lexical candidates for sentence #1:

more:= $\mathbf{NP(X)/N(X)} : \mathbf{more}'_{((e,t),e)}$

$N(\text{more}) : \mathit{dogs}'_{(e,t)}$

doggies:= $\mathbf{NP(X) \setminus N(X)} : \mathit{more}'_{((e,t),e)}$

$\mathbf{N(\text{doggies})} : \mathbf{dogs}'_{(e,t)}$

more doggies:= $\mathbf{NP(\text{more doggies})} : (\mathit{more}' \mathit{dogs}')_e$

- **She might get it wrong**, starting to use “doggies” to mean “more”. But she soon corrects in the light of further evidence.
- **Where *more' dog'* came from is a different question—see Quine (1960).**

5

Computational Accounts: Zettlemoyer and Collins

- There is no reason to separate the two processes of associating meaning and syntactic type. Zettlemoyer and Collins (UAI 2005) combine the two in a single pass CCG induction algorithm which seems both simpler and more general.
- Crucially, their algorithm allows **any contiguous substring** of the sentence to be a lexical item, so that for the given logical form, the learner has to search the cross-product of the substring powerset of the string with the set of pairs of legal categories with substructure powerset of the logical form, as in the example (1) for categories that yield combinatory derivations that yield the correct logical form.
- Learning is via a log-linear model using lexical entries as features and gradient descent on their weights, iterating over successive sentences of a corpus of sentence-logical form pairs.

7

Computational Accounts

- Siskind (1995, 1996), Villavicencio (2002), and Zettlemoyer and Collins (2005) offer computational models of this process.
- Both theories make strong assumptions about the association of words with elements of logical form.
- Both make strong assumptions about universally available parametrically specified rule- or category- types, the latter in the form of a type hierarchy
- Both deal with noise and homonymy probabilistically.
- Both do the learning in two stages:
 - Association of logical forms with words.
 - Induction of phrase structure rules (Siskind) or directional CCG categories (Villavicencio).
 - The first is reminiscent of **alignment** in MT data. The second is reminiscent of **learning recursive phrasal MT rules** by Chiang (2005).

6

Zettlemoyer and Collins (Contd.)

- The algorithm as presented in 2005 learns only a very small rather unambiguous fragment of English, hand-labeled with uniquely identified database queries as logical forms, and an English specific inventory of possible syntactic category types in lieu of Universal Grammar.
- However, Siskind's and Villavicencio's results already tell us that the algorithm should work with multiple candidate logical forms.
- Similarly, their results show that a universal set of category types can be used without overwhelming the learner.
- **All of these models depend on availability to the learner of short sentences** paired with logical forms, since complexity is determined by a cross-product of powersets both of which are exponential in sentence length.
- A number of techniques are available to make search efficient including **use of a head-dependency parsing model**.

8

Learning with a Generative model: Preliminaries

- We can view an estimator for an increment to the cumulative expected frequency (??) based on observing the n th sentence as the sum of two weighted components, an “a posteriori” component, stemming from what we have already learned, and an “prior” component defined by all possibilities allowed by universal grammar—thus:

$$(4) \Delta fexp = \lambda \Delta fexp_{prior} + (1 - \lambda) \Delta fexp_{post}$$

- $fexp_{prior}$ is the expected frequency based on the present sentence and the possibilities of universal grammar alone. It is defined as follows, where P is whatever distribution UG imposes on the derivations for the present sentence. For simplicity we will assume it is a uniform distribution, so that (??) reduces to the following, where $|D|$ is the number of derivations:

$$(5) fexp_{prior}(p) = \frac{f(p,d)}{|D|}$$

- $fexp_{post}$ for a given interpretation i for sentence s is defined as follows, where

9

Learning with a Generative model: EM Algorithm

- Such a model, including the implicit changes in the λ s over time, can be learned using the following incremental variant of the semi-supervised inside-outside (EM) algorithm (Pereira and Schabes 1992; Neal and Hinton 1999).
- Every new sentence s_n provides a set D_n of derivations parallel to (2), which defines the following:
 - a. A (possibly empty) set of previously unseen productions involved in some derivation in D_i , including those involving novel lexical entries, that must be added to the model with cumulative $fexp$ temporarily initialized to zero.
 - b. (E-step): The set of all productions including those in a, whose cumulative $fexp$ must be multiplied by $n - 1$, incremented by $fexp_{prior}$, and divided by n .
 - c. (M-step): A further increment of $\frac{fexp_{post} - fexp_{prior}}{n}$ (which may be negative) to the cumulative $fexp$ for all productions involved in some derivation in D_i .

11

P is the model estimated so far.

$$(6) fexp_{post}(p) = \sum_{i \in I} P(i|s) \sum_{d \in D} P(d|s, i) \cdot f(p, d)$$

10

- Step b defines new values for the conditional probabilities for the rules in question, defining an intermediate model for calculating the a posteriori probabilities in step c. The further update c to the model defines the expected frequencies for the next cycle. The lexical probabilities for the relevant words in the lexicon given the new sentence can then be calculated using the model and definition (??), where $P(D|I, S)$ is the product of the probabilities of the productions it involves.

12

Learning with a Generative model

- In the case of our running example and the first sentence “More doggies”, this process can be viewed as follows:

(7) *The Child’s First Parsing Model:*

Rule	$f_{exp}(n-1)$	$\frac{(n-1)f_{exp}(n-1)+f_{exp}prior}{n}$	$f_{exp}(n)$
r0. $START \rightarrow NP(X)$	0	1.0	1.0
r1. $NP(X) : fa \rightarrow NP(X)/N(X) : f \quad N(X) : a$	0	0.3	0.3
r2. $NP(X) : fa \rightarrow N(X) : a \quad NP(X) \setminus N(X) : f$	0	0.3	0.3
r3. $NP(more)/N(more) : more' \rightarrow more$	0	0.3	0.3
r4. $NP(doggies) \setminus N(doggies) : more' \rightarrow doggies$	0	0.3	0.3
r5. $N(doggies) : dogs' \rightarrow doggies$	0	0.3	0.3
r6. $N(more) : dogs' \rightarrow more$	0	0.3	0.3
r7. $NP(more doggies) : more' dogs' \rightarrow more doggies$	0	0.3	0.3

13

Learning with a Generative model

- Thus, by (??), we have the following updated probabilistic lexicon:

(9) *The Child’s First Lexicon:*

ϕ	σ, μ	f_{exp}	$P_{lex}(\sigma, \mu \phi)$	$P_{lex}(\phi \mu)$
more:=	NP/N : $more'_{((e,t),e)}$	0.3	0.5	0.5
	N : $dogs'_{(e,t)}$	0.3	0.5	0.5
doggies:=	$NP \setminus N$: $more'_{((e,t),e)}$	0.3	0.5	0.5
	N : $dogs'_{(e,t)}$	0.3	0.5	0.5
more doggies:=	NP : $(more' dogs')_e$	0.3	0.3	0.3

- Since the word counts and conditional probabilities for “more” and “doggies” with them meaning $more'_{((e,t),e)}$ are all equal at this stage, the child may well make errors of overgeneralization, using some approximation to “doggies” to mean “more”.^a

^aThe example is based on an attested case of this particular error (Cathy Urwin, p.c.).

15

Learning with a Generative model

- On the basis of the intermediate value $\frac{(0)f_{exp}(0)+f_{exp}prior}{1}$, the relative conditional probabilities $P(D|I, S)$ of the three derivations (2) are as follows:

$$(8) \text{ a } P(A|I, S) = P(START \rightarrow NP(doggies) : dogs' | START) \times P(r1 | NP(doggies) : more' dogs', r1, 1) \times P(NP(more)/N(more) : more | NP(doggies) : more' dogs', r1, 1) \times P(N(doggies) : dogs' | NP(doggies) : more' dogs', r1, 2) \times P(more | NP(more)/N(more) : more') \times P(doggies | N(doggies) : dogs) = \frac{1 \times 0.3 \times 1.0 \times 1.0 \times 1.0 \times 1.0}{\sum_d P(d|I, S)} = P(B|I, S) = P(C|I, S) = 0.3$$

(NOTE: Actually we need to normalize for length of derivation)

- Thus, the further increment (c) due to posterior expected frequency an be calculated, to determine $f_{exp}(I)$. In the case of this first sentence, $f_{exp}post = f_{exp}prior$, so that $f_{exp}post - f_{exp}prior = 0$, and the implicit λ is 1 for all rules.

14

- However, even on the basis of this very underspecified lexicon, the child will not overgenerate “*doggies more”.^b

^bIt follows that overgeneralizations by the child like “Allgone doggies” must arise from processes of lexical generalization of the category for “more” to a meaning *allgone'* of the same semantic type as *more'*.

16

Learning with a Generative model

- Let us suppose that the second utterance the child hears is “More cookies”. There are again three derivations parallel to (2).

(10) The Child’s Parsing Model #2:

Rule	$f_{exp}(n-1)$	$\frac{(n-1)f_{exp}(n-1)+f_{exp}prior}{n}$	$f_{exp}(n)$
r0. $START \rightarrow NP(X)$	1.0	1.0	1.0
r1. $NP(X) : fa \rightarrow NP(X)/N(X) : f \ N(X) : a$	0.3	0.3	0.23
r2. $NP(X) : fa \rightarrow N(X) : a \ NP(X) \setminus N(X) : f$	0.3	0.3	0.20
r3. $NP(X)/N(X) : more' \rightarrow more$	0.3	0.3	0.23
r4. $NP(X) \setminus N(X) : more' \rightarrow doggies$	0.3	0.16	0.16
r5. $N(doggies) : dogs' \rightarrow doggies$	0.3	0.16	0.16
r6. $N(more) : dogs' \rightarrow more$	0.3	0.16	0.16
r7. $NP(more doggies) : more' dog' \rightarrow more doggies$	0.3	0.16	0.16
r8. $NP(more cookies) : more' cookies' \rightarrow more cookies$	0	0.16	0.40
r9. $NP(X) \setminus N(X) : more' \rightarrow cookies$	0	0.16	0.03
r10. $N(cookies) : cookies' \rightarrow cookies$	0	0.16	0.0513
r11. $N(more) : cookies' \rightarrow more$	0	0.16	0.03

17

calculated, to determine $f_{exp}(2)$. In the case of the second and all subsequent sentences, $f_{exp}post \neq f_{exp}prior$, so that $f_{exp}post - f_{exp}prior \neq 0$, and the implicit $\lambda < 1$ for all rules.

19

Learning with a Generative model

- On the basis of the intermediate value $\frac{(1)f_{exp}(1)+f_{exp}prior}{2}$, the length-weighted relative conditional probabilities $P(d|I,S)$ of the three derivations for “More cookies” parallel to (2) are as follows:

$$(11) \text{ a' } P(A|I,S) = P(START \rightarrow NP(cookies) : more' cookies' | START) \times P(r1|NP(cookies) : more' cookies') \times P(NP(cookies)/N(cookies)|NP(cookies) : more' cookies, r1, I) \times P(N(cookies) : cookies'|NP(cookies) : more' cookies', r1, 2) \times P(more|NP(cookies)/N(cookies) : more') \times P(cookies|N(cookies) : cookies') = \frac{1 \times 0.3 \times 1 \times 1 \times 0.3 \times 0.16}{\sum_d P(d|I,S)} \times \frac{6}{14} = 0.13$$

$$\text{ b' } P(B|I,S) = P(START \rightarrow NP(more) : more' cookies' | START) \times P(r1|NP(more) : more' cookies') \times P(NP(more) \setminus N(more) : more' | NP(more) : more' cookies, r1, 2) \times P(N(more) : cookies'|NP(more) : more' cookies', r1, 1) \times P(cookies|NP(more) \setminus N(more) : more') \times P(more|N(more) : cookies') = \frac{1 \times 0.3 \times 1 \times 1 \times 0.16 \times 0.16}{\sum_d P(d|I,S)} \times \frac{6}{14} = 0.06$$

$$\text{ c' } P(C|I,S) = P(START \rightarrow NP(more cookies) : more' cookies' | START) \times P(more cookies|NP(more cookies) : more' cookies') = \frac{1 \times 0.3 \times 1 \times 1 \times 0.16 \times 0.16}{\sum_d P(d|I,S)} \times \frac{2}{14} = 0.80$$

- Thus, the further increment (c) due to posterior expected frequency can be

18

Learning with a Generative model

- Thus, by (??), we have the following updated probabilistic lexicon:

(12) The Child’s Lexicon #2:

ϕ	σ, μ	$f_{exp}lex(n)$	$P(\sigma, \mu \phi)$	$P(\phi \mu)$
more:=	NP/N : more' _{((e,t),e)}	0.23	0.538	0.538
	$N : dogs'$ _(e,t)	0.16	0.385	0.385
	$N : cookies'$ _(e,t)	0.03	0.077	0.077
doggies:=	$NP \setminus N : more'$ _{((e,t),e)}	0.16	0.5	0.385
	N : dogs' _(e,t)	0.16	0.5	0.50
cookies:=	$NP \setminus N : more'$ _{((e,t),e)}	0.03	0.394	0.077
	N : cookies' _(e,t)	0.0513	0.606	0.606
more doggies:=	$NP : (more' dogs')_e$	0.16	1.0	1.0
more cookies:=	$NP : (more' cookies')_e$	0.40	1.0	1.0

- Notice that the expected frequencies in this table are not quite the same as

20

those that would be obtained by recomputing f_{exp} over the entire corpus. If we did that, then we would realize that the expected frequency of $more := N : dogs'$ is actually 0.06, and that of $more := NP/N : more'$, 0.26, among other differences.

- However, this approximation will become more exact as more sentences are analyzed, and it is worth tolerating it in order not to have to make the implausible assumption that the child keeps a corpus of all analyses of all sentences it has ever encountered and recomputes the model from scratch at each iteration.
- Despite the inexactness of the lexical expected frequencies, the probability that the child will correctly say “more” when they mean $more'$ is already greater than that of spurious candidates like “doggies” or “cookies.”

21

References

- Chiang, David, 2005. “A Hierarchical Phrase-Based model for Statistical Machine Translation.” In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor MI: ACL, 263–270.
- McWhinnie, Brian, 2005. “Item Based Constructions and the Logical Problem.” In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*. CoNLL-9. New Brunswick: ACL, 53–68.
- Neal, Radford and Hinton, Geoffrey, 1999. “A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants.” In Michael Jordan (ed.), *Learning in Graphical Models*, Cambridge, MA: MIT Press. 355–368.
- Pereira, Fernando and Schabes, Yves, 1992. “Inside-Outside Reestimation from Partially Bracketed Corpora.” In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. ACL, 128–135.
- Quine, Willard van Ormond, 1960. *Word and Object*. Cambridge MA: MIT Press.

23

Zettlemoyer and Collins (Contd.)

- Because it allows multiword elements (MWE) to be lexical entries, it avoids the problem that two words which consistently collocate, like *want* and *to* fails to reveal which of them means $want'$ and which means to' . They can be learned as a single item *want to*
- So can idioms/MWEs like “buy the farm,” and “take advantage of”
- As with Siskind’s version lexical items can have complex meanings—corresponding for example to causatives, whose availability may differ (*swim across* vs. *traverser à la nâge*) across languages.
- No notion of “triggers” distinct from reasonably short string-meaning pairs is necessary.
- It is possible to use the statistics of the lexicon itself to implicitly represent “parameters” such as verb-finality, via incrementally adjusted prior probabilities on the members of the set of universally available category types.

22

- Siskind, Jeffrey, 1995. “Grounding Language in Perception.” *Artificial Intelligence Review* 8:371–391.
- Siskind, Jeffrey, 1996. “A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings.” *Cognition* 61:39–91.
- Villavicencio, Aline, 2002. *The Acquisition of a Unification-Based Generalised Categorical Grammar*. Ph.D. thesis, University of Cambridge.
- Zettlemoyer, Luke and Collins, Michael, 2005. “Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorical Grammars.” In *Proceedings of the 21st Conference on Uncertainty in AI (UAI)*. ACL, 658–666.

24