Cognitive Modeling

Lecture 15: Probabilistic Models of Syntactic Processing

Sharon Goldwater

School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

March 1, 2010

Reading: Jurafsky (1996).

Jurafsky (1996)

Covers a lot of ground: a unified probabilistic account of much previous work, explaining frequency and context effects.

- Lexical access (word recognition)
- Idiom/phrase access
- Syntactic processing (access and disambiguation)

We'll focus on syntactic processing.

- Model shows how augmenting parallel parser with probabilities can explain garden paths and disambiguation.
- By analogy with lexical access, Jurafsky then argues for parallel over serial architecture.

Disambiguation

Main assumptions of Jurafsky (1996):

- Observed preferences in interpretation of ambiguous sentences reflect *probabilities* of different syntactic structures.
- Garden path effects are merely extreme cases of processing preferences.

Examples from several types of ambiguity:

- Lexical category ambiguity
- Attachment ambiguity
- Main clause vs. reduced relative clause ambiguity

Lexical category ambiguity

Ambiguity resolved without trouble (*fires* = N or V):

(1) a. The warehouse fires destroyed all the buildings.
b. The warehouse fires a dozen employees each year.

Ambiguity leads to garden path (*complex*= N or Adj, *houses*= N or V, etc.):

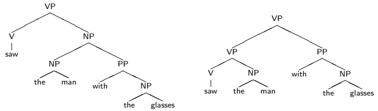
(2) a. #The complex houses married and single students.
b. #The old man the boats.

Note: # means garden path.

Attachment ambiguity

Prepositional phrase can attach to NP or VP.

(3) I saw the man with the glasses.



(4) #The landlord painted the walls with cracks.

Subcategorization frames

Attachment preferences vary between verbs (Ford et al. 1982):

(5) The women discussed the dogs on the beach.
a. The women discussed the dogs that were on the beach. (90%)
b. The women discussed the dogs while on the beach. (10%)

(6) The women kept the dogs on the beach.
a. The women kept the dogs that were on the beach. (5%)
b. The women kept them (the dogs) on the beach. (95%)

The arguments required by a verb are its *subcategorization frame* or *valence*. Different valence preferences create different attachment preferences.

Main clause vs. reduced relative clause

Reduced relative clause: *that*-clause without the *that*.

(7) a. #The horse raced past the barn fell.
b. The horse found in the woods died.

Another case of different subcategorization preferences:

- X raced >> X raced Y
- X found Y >> X found

Serial parsing

- if multiple rules can apply, choose one based on a selection rule;
- if parse fails, backtrack to choice point and reparse;
- example selection rule: minimal attachment (choose the tree with the fewest nodes).
- garden path means the wrong tree was selected at a choice point;
- backtracking occurs, causes increased processing times.

Parallel parsing

- if multiple rules can apply, pursue all possibilities in parallel;
- if any parse fails, discard it;
- problem: number of parse trees can grow exponentially.
- solution: only pursue a limited number of possibilities (*bounded parallelism*).
- garden path means correct tree was pruned from search space;
- backtracking occurs, causes increased processing times.

A probabilistic parallel parser

Jurafsky (1996) adopts probabilistic parsing techniques from computational linguistics in a parallel parsing model.

- Each full or partial parse is assigned a probability.
- Parses are pruned from the search space if their probability is a factor of α below the most probable parse (*beam search*).
- Other pruning methods are possible, e.g., maintain a fixed number of parses at all times.

How are parse probabilities determined?

Computing parse probabilities

Jurafsky (1996) focuses on two sources of information:

- *Construction probabilities*: probability of syntactic tree.
- *Valence probabilities*: probability of particular syntactic categories as arguments for specific verbs.

Assumes that construction probabilities and valence probabilities

- are independent, so

$$P(parse) = P(constructions) * P(subcat\ frames)$$

- can be estimated from a large *treebank* using relative frequencies.

(Note: parts of the paper use Construction Grammar formalism; this is slightly different from the *construction* in *construction probabilities*.)

Probabilistic Context-free Grammars

P(constructions) is computed as $P_{pcfg}(parse)$.

Example (Manning and Schütze 1999)

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → V NP	0.7	NP → ears	0.18
VP → VP PP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

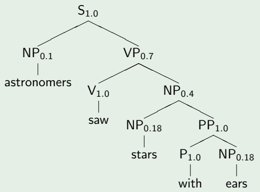
- The rule A → B C with probability p means

$$P(\textit{left-hand side expands as B C} \mid \textit{left-hand side is A}) = p$$

- so, probabilities of all rules with the same LHS sum to one;
- $P_{pcfg}(parse) = \prod P_{pcfg}(rule_i)$ of all rules applied in the parse.

Probabilistic Context-free Grammars

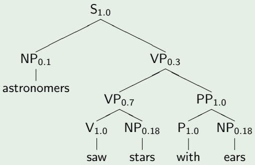
Example (Manning and Schütze 1999)



$P(t_1) = 1.0 \cdot 0.1 \cdot 0.7 \cdot 1.0 \cdot 0.4 \cdot 0.18 \cdot 1.0 \cdot 1.0 \cdot 0.18 = 0.0009072$

Probabilistic Context-free Grammars

Example (Manning and Schütze 1999)



$P(t_2) = 1.0 \cdot 0.1 \cdot 0.3 \cdot 0.7 \cdot 1.0 \cdot 0.18 \cdot 1.0 \cdot 1.0 \cdot 0.18 = 0.0006804$

t_1 more probable than t_2.

Valence Probabilities

Subcategorization frames of the verb *keep:*

NP AP	keep the prices reasonable
NP VP	keep his foes guessing
NP VP	keep their eyes peeled
NP PRT	keep the people in
NP PP	keep his nerves from jangling

Valence probabilities tell us how likely each of these frames is.

Valence Probabilities

Like PCFG probabilities, valence probabilities are estimated from a treebank.

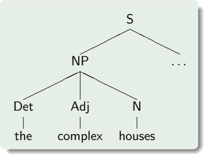
Example

discuss	⟨NP PP⟩	.24
	⟨NP⟩	.76
keep	⟨NP XP[pred +]⟩	.81
	⟨NP⟩	.19

Modeling Garden Path Effects

Garden path caused by construction probabilities:

S → NP ...	0.92	N → houses	0.00055
NP → Det Adj N	0.28	Adj → complex	0.00086
Det → the	0.71		

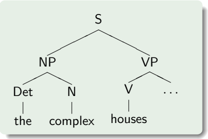


$p(t_1) = 8.5 \cdot 10^{-8}$ (preferred)

Modeling Garden Path Effects

Garden path caused by construction probabilities:

NP → Det N	0.63	V → houses	0.000052
S → [NP $_{VP}$[V ...	0.48	Det → the	0.71
N → complex	0.000029		

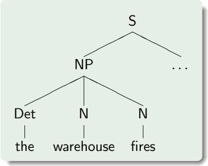


$p(t_2) = 3.2 \cdot 10^{-10}$ (grossly dispreferred)

Modeling Disambiguation

Disambiguation using construction probabilities, no garden path:

S → NP ...	0.92	N → fires	0.00017
NP → Det N N	0.28		

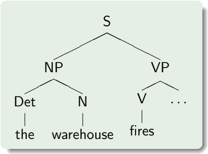


$p(t_1) = 4.2 \cdot 10^{-5}$ (preferred)

Modeling Disambiguation

Disambiguation using construction probabilities, no garden path:

NP → Det N   0.63   V → fires   0.000036
S → [NP $_{VP}$[V …   0.48

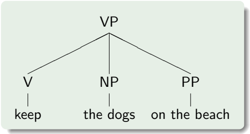


$p(t_2) = 1.1 \cdot 10^{-5}$ (mildly dispreferred)

Modeling Valence Preferences

Disambiguation using valence probabilities, no garden path:

$p(\text{keep}, \langle \text{NP XP[pred +]} \rangle) = 0.81$
VP → V NP XP   0.15

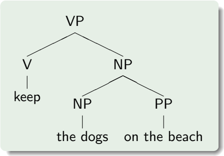


$p(t_1) = 0.15 \cdot 0.81 = 0.12$ (preferred)

Modeling Valence Preferences

Disambiguation using valence probabilities, no garden path:

$p(\text{keep}, \langle \text{NP} \rangle) = 0.19$ VP → V NP   0.39
NP → NP XP   0.14

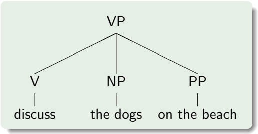


$p(t_2) = 0.19 \cdot 0.39 \cdot 0.14 = 0.01$ (mildly dispreferred)

Modeling Valence Preferences

Disambiguation using valence probabilities, no garden path:

$p(\text{discuss}, \langle \text{NP PP} \rangle) = 0.24$
VP → V NP XP   0.15

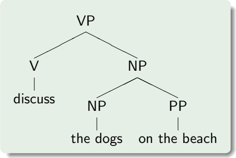


$p(t_1) = 0.15 \cdot 0.24 = 0.036$ (mildly dispreferred)

Modeling Valence Preferences

Disambiguation using valence probabilities, no garden path:

$p(\text{discuss}, \langle\text{NP}\rangle) = 0.76$ VP → V NP   0.39
NP → NP XP   0.14

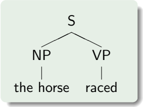


$p(t_2) = 0.76 \cdot 0.39 \cdot 0.14 = 0.041$ (preferred)

Combining valence and construction probabilities

Garden path caused by construction probabilities and valence probabilities:

$p(\text{race}, \langle\text{NP}\rangle) = 0.92$

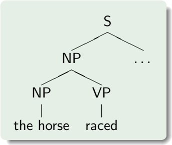


$p(t_1) = 0.92$ (preferred)

Combining valence and construction probabilities

Garden path caused by construction probabilities and valence probabilities:

$p(\text{race}, \langle\text{NP NP}\rangle) = 0.08$
NP → NP XP   0.14

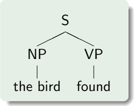


$p(t_2) = 0.0112$ (grossly dispreferred)

Combining valence and construction probabilities

Disambiguation using construction probabilities and valence probabilities, no garden path:

$p(\text{find}, \langle\text{NP}\rangle) = 0.38$



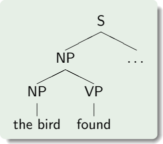
$p(t_1) = 0.38$ (preferred)

Combining valence and construction probabilities

Disambiguation using construction probabilities and valence probabilities, no garden path:

$p(\text{find}, \langle \text{NP NP} \rangle) = 0.62$

NP $\rightarrow$ NP XP 0.14



$p(t_2) = 0.0868$ (mildly dispreferred)

Setting the Beam Width

Crucial assumption: if the relative probability of a tree falls below a certain value, then it will be pruned.

sentence	probability ratio
the complex houses ...	267:1
the horse raced ...	82:1
the warehouse fires ...	3.8:1
the bird found ...	3.7:1

Assumption: a garden path occurs if the probability ratio is higher than 5:1.

Open Issues

- Incrementality: Can we make more fine-grained predictions of the time course of ambiguity resolution?
- Coverage: Jurafsky used hand-crafted examples. Can we use a probabilistic parser that is trained on a real corpus?
- Crosslinguistics: does this model work for languages other than English?

Summary

- Different types of garden paths: main clause/reduced relative; attachment ambiguity; lexical category;
- rating studies provide evidence for subcat frame preferences;
- modeling assumption:
  - parser with bounded parallelism;
  - pruning of improbable analyses (beam search);
  - independent combination of PCFG and valence probabilities;
- Model accounts for human parse preferences in several well-known examples.
- beam width: ratio of the probability of the preferred analysis to the dispreferred analysis; needs to be determined empirically.

References

Ford, Marilyn, Joan Bresnan, and Ronald M. Kaplan. 1982. A competence-based theory of syntactic closure. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, MA, pages 727–796.

Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2):137–194.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.