Connectionist-Inspired Incremental PCFG Parsing

Marten van Schijndel[a], Andy Exley[b], William Schuler[a]
[a]Dept Linguistics, The Ohio State University
[b]Dept Computer Science and Engineering, University of Minnesota

June 7, 2012

Introduction

Goals and Motivation

Create a cognitively-motivated parser

- [Schuler, 2009] outlines a cognitively-motivated parser, which requires book-keeping nodes built in to work with PCFGs (engineering fix).
- We'd like to be able to strip out elements included solely for engineering.

Background

Why PCFGs? [Jurafsky, 1996]

- Simple
- Widespread use, community understanding
- Easily integrated with other technologies
- Latent variable training procedures easily obtained [Petrov et al., 2006]
- Tractable recognition $\mathcal{O}(n^3)$

Problems with CKY

- Not incremental $\mathcal{O}(n^3)$
- In certain applications, word/phrase breaks not certain (ASR, MT, etc)

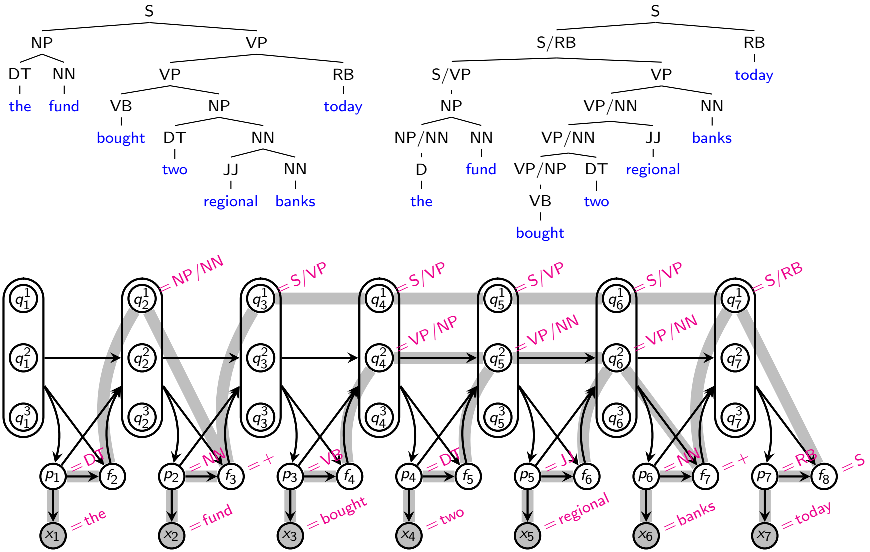
Background

Why Incremental?

- Operates on incomplete information
- Can make use of information about recent content/structure
- $\mathcal{O}(n)$
- Streaming task

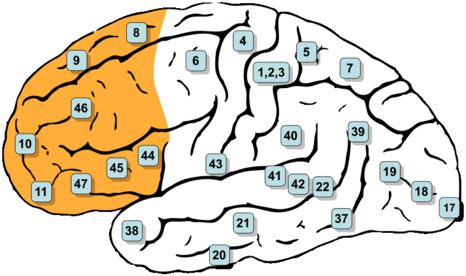
Must operate on a beam to efficiently stream

The Setup



S
NP
VP
DT
NN
the
fund
VP
RB
today
VB
NP
bought
DT
NN
two
JJ
NN
regional
banks
S
S/RB
RB
today
S/VP
VP
NP
VP/NN
NN
NP/NN
NN
VP/NN
JJ
banks
D
fund
VP/NP
DT
regional
the
VB
two
bought
=NP/NN
=S/VP
=S/VP
=S/VP
=S/VP
=S/RB
=VP/NP
=VP/NN
=VP/NN
=DT
=NN
=+
=VB
=DT
=JJ
=NN
=+
=RB
=S
=the
=fund
=bought
=two
=regional
=banks
=today

Neural Motivation



- Corresponding structure seen in C-R axis of DL-PFC (proximal to Broca's) [Petrides, 1987, Botvinick, 2007]

Cognitive Motivation

- Can define graph-theory connected components (sub-graphs) of a semantic dependency graph (of 'concepts' [Kintsch, 1988] or discourse referents)
- F-node = create new independent connected component linked via an episodic trace [Sederberg et al., 2008] to previous connected component
- Connected components act as 'chunks' [Miller, 1956]

Design Motivations

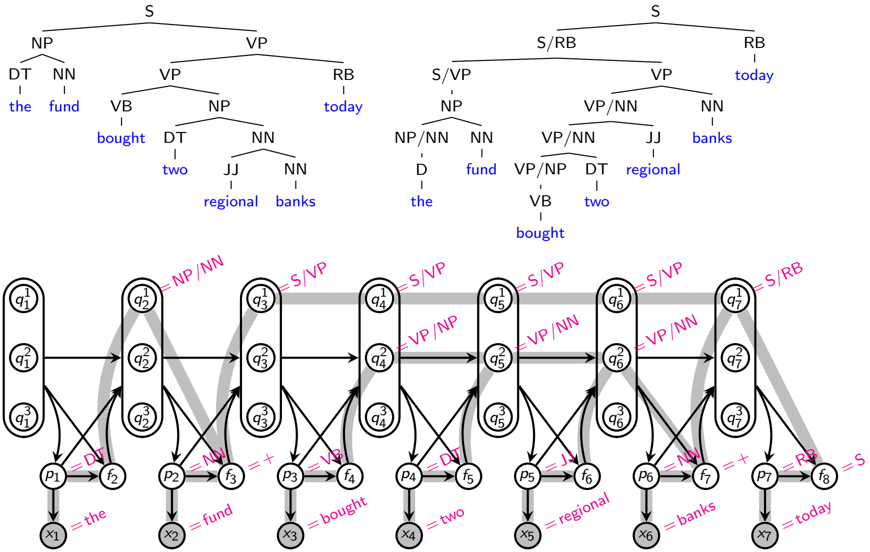
Schuler (2009) based on:

- HHMM [Murphy and Paskin, 2001] but too general (next slide)
- 4 layers [Cowan, 2001]

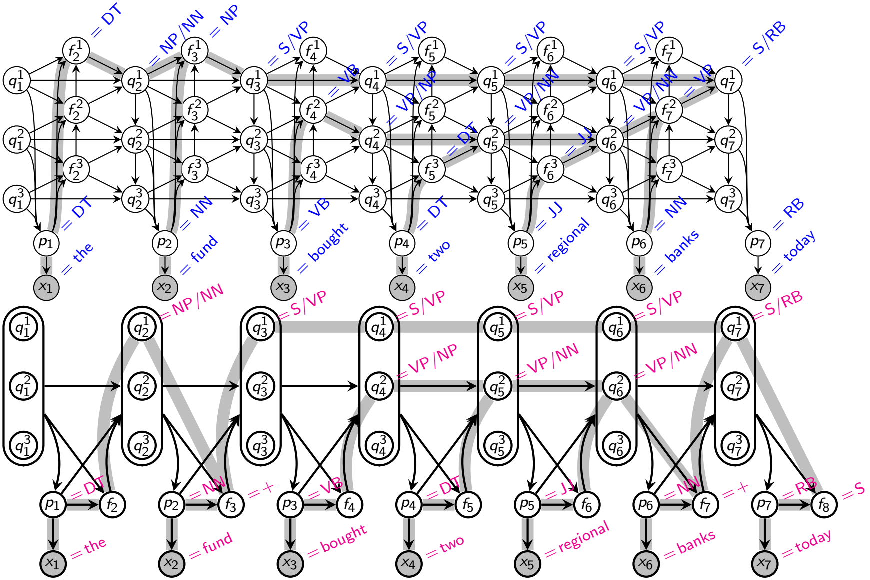
Serial recall chunking [Miller, 1956] seems to be different from language chunking or chunking with distractions [Cowan, 2001].

[Schuler et al., 2010] found 4 layers yielded >99.9% coverage of WSJ.

Single Expansion, Single Reduction



The Model



= DT
= NP/NN
= NP
= S/VP
= S/VP
= S/VP
= S/VP
= S/RB
= VB
= VP/NP
= VP/NN
= VP/NN
= DT
= JJ
= DT
= NN
= VB
= DT
= JJ
= NN
= RB
= the
= fund
= bought
= two
= regional
= banks
= today
= NP/NN
= S/VP
= S/VP
= S/VP
= S/VP
= S/RB
= VP/NP
= VP/NN
= VP/NN
= DT
= NN
= +
= VB
= DT
= JJ
= NN
= +
= RB
= S
= the
= fund
= bought
= two
= regional
= banks
= today

Tree Training

Split-Merge Berkeley Grammar Trainer
[Petrov et al., 2006]

- Input: TB-annotated sentences
  (S (ADVP happily) (NP-SUBJ John). . .)

Tree Training

Split-Merge Berkeley Grammar Trainer
[Petrov et al., 2006]

- Input: TB-annotated sentences
  (S (ADVP happily) (NP-SUBJ John)...)
- EM classification performed over a given number of split-merge cycles
- Output: Subcat-Annotated PCFG
  (S^g_10 -> ADVP^g_21 NP^g_4 1.462527E-18)

Profit:

- More specialized and informative PCFG

Cost:

- Training time
- Increased size of grammar

Through the Crucible

Testing Methodology

Internal Testing

- Timing Comparisons [Hidden State Factoring]

External Testing

- Roark (2001) Parser [Incremental]
- Petrov and Klein (2007) Parser [CKY Chart Parser]

Paydirt

Accuracy Results

System	R	P	F
Schuler et al. 2008/2010	83.4	83.7	83.5
Roark 2001	86.6	86.5	86.5
Schuler 2009* (2000)	87.9	87.8	87.8
van Schijndel et al (250)	85.6	87.1	86.3
van Schijndel et al (500)	86.8	87.4	87.1
van Schijndel et al (1000)	87.4	87.6	87.5
van Schijndel et al (2000)	87.9	87.8	87.8
van Schijndel et al (5000)	87.9	87.8	87.8
Petrov Klein (Binary)	88.1	87.8	88.0
Petrov Klein (+Unary)	88.3	88.6	88.5

*Without grammar trainer, Schuler 2009 (2000) F-Score = 75.06.

Paydirt

Timing Results

System	Sec/Sent
Schuler 2009	74
Current Model	12

Table : Speed comparison using a beam-width of 500 elements

Digging Deeper

Future Work

- Incremental Dependency Parsing (including Unbounded)
- Incremental Semantic Role Labelling
- Interactive associative memory access
- Coreference resolution

Questions?

Thanks!

Overtime

More slides!

Paydirt

Full Accuracy Results

System	R	P	F
Schuler et al. 2008/2010	83.4	83.7	83.5
Roark 2001	86.6	86.5	86.5
Schuler 2009 (2000)	87.9	87.8	87.8
van Schijndel et al (50)	75.9	84.6	80.0
van Schijndel et al (100)	81.7	85.6	83.6
van Schijndel et al (250)	85.6	87.1	86.3
van Schijndel et al (500)	86.8	87.4	87.1
van Schijndel et al (1000)	87.4	87.6	87.5
van Schijndel et al (1500)	87.6	87.7	87.7
van Schijndel et al (2000)	87.9	87.8	87.8
van Schijndel et al (5000)	87.9	87.8	87.8
Petrov Klein (Binary)	88.1	87.8	88.0
Petrov Klein (+Unary)	88.3	88.6	88.5

How does it work?

Theory/Equation time

Most likely sequence

$$\hat{q}^{1..D}_{1..T} \stackrel{\text{def}}{=} \underset{q^{1..D}_{1..T}}{\operatorname{argmax}} \prod_{t=1}^{T} \mathsf{P}_{\theta_Q}(q^{1..D}_t \mid q^{1..D}_{t-1}\ p_{t-1}) \cdot \mathsf{P}_{\theta_{P,d'}}(p_t \mid b^{d'}_t) \cdot \mathsf{P}_{\theta_X}(x_t \mid p_t) \quad (1)$$

where d' is the lowest non-empty q^d_t

How does it work?

Theory/Equation time

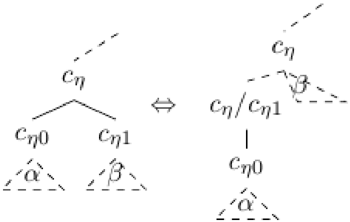
Right-Corner: Single expansion, Single reduction

E-R+, E-R-, E+R+, E+R-

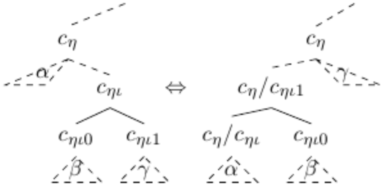
θ_Q

$$
\begin{aligned}
&\mathsf{P}_{\theta_Q}(q_t^{1..D} \mid q_{t-1}^{1..D}\; p_{t-1}) \\
&\quad \stackrel{\text{def}}{=} \mathsf{P}_{\theta_F}(\text{`0'} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot \mathsf{P}_{\theta_{A,d'}}(\text{`–'} \mid b_{t-1}^{d'-1}\; a_{t-1}^{d'}) \cdot [\![a_t^{d'-1} = a_{t-1}^{d'-1}]\!] \cdot \mathsf{P}_{\theta_{B,d'-1}}(b_t^{d'-1} \mid b_{t-1}^{d'-1}\; a_{t-1}^{d'}) \\
&\qquad \cdot [\![q_t^{1..d'-2} = q_{t-1}^{1..d'-2}]\!] \cdot [\![q_t^{d'..D} = \text{`–'}]\!] \\
&\quad + \mathsf{P}_{\theta_F}(\text{`0'} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot \mathsf{P}_{\theta_{A,d'}}(a_t^{d'} \mid b_{t-1}^{d'-1}\; a_{t-1}^{d'}) \cdot \mathsf{P}_{\theta_{B,d'}}(b_t^{d'} \mid a_t^{d'}\; a_{t-1}^{d'+1}) \\
&\qquad \cdot [\![q_t^{1..d'-1} = q_{t-1}^{1..d'-1}]\!] \cdot [\![q_t^{d'+1..D} = \text{`–'}]\!] \\
&\quad + \mathsf{P}_{\theta_F}(\text{`1'} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot \mathsf{P}_{\theta_{A,d'}}(\text{`–'} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot [\![a_t^{d'} = a_{t-1}^{d'}]\!] \cdot \mathsf{P}_{\theta_{B,d'}}(b_t^{d'} \mid b_{t-1}^{d'}\; p_{t-1}) \\
&\qquad \cdot [\![q_t^{1..d'-1} = q_{t-1}^{1..d'-1}]\!] \cdot [\![q_t^{d'+1..D} = \text{`–'}]\!] \\
&\quad + \mathsf{P}_{\theta_F}(\text{`1'} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot \mathsf{P}_{\theta_{A,d'}}(a_t^{d'+1} \mid b_{t-1}^{d'}\; p_{t-1}) \cdot \mathsf{P}_{\theta_{B,d'}}(b_t^{d'+1} \mid a_t^{d'+1}\; p_{t-1}) \\
&\qquad \cdot [\![q_t^{1..d'} = q_{t-1}^{1..d'}]\!] \cdot [\![q_t^{d'+2..D} = \text{`–'}]\!]
\end{aligned} \tag{2}
$$

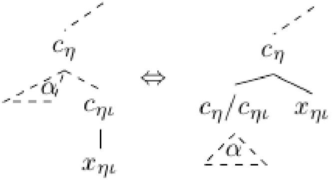
The right-corner transform (tree)



(1)



(2)



(3)

The right-corner transform (grammar)

$$\frac{c_\eta \rightarrow c_{\eta 0}\ c_{\eta 1} \in G}{c_\eta / c_{\eta 1} \rightarrow c_{\eta 0} \in G'} \tag{1}$$

$$\frac{c_{\eta\iota} \rightarrow c_{\eta\iota 0}\ c_{\eta\iota 1} \in G, \quad c_\eta \in C}{c_\eta / c_{\eta\iota 1} \rightarrow c_\eta / c_{\eta\iota}\ c_{\eta\iota 0} \in G'} \tag{2}$$

$$\frac{c_{\eta\iota} \rightarrow x_{\eta\iota} \in G, \quad c_\eta \in C}{c_\eta \rightarrow c_\eta / c_{\eta\iota}\ c_{\eta\iota} \in G'} \tag{3}$$

Bibliography I

Botvinick, M. (2007).
Multilevel structure in behavior and in the brain: a computational model of fuster's hierarchy.
Philosophical Transactions of the Royal Society, Series B: Biological Sciences, 362:1615–1626.

Cowan, N. (2001).
The magical number 4 in short-term memory: A reconsideration of mental storage capacity.
Behavioral and Brain Sciences, 24:87–185.

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

Bibliography II

Kintsch, W. (1988).
The role of knowledge in discourse comprehension: A construction-integration model.
Psychological review, 95(2):163–182.

Miller, G. A. (1956).
The magical number seven, plus or minus two: Some limits on our capacity for processing information.
Psychological Review, 63:81–97.

Murphy, K. P. and Paskin, M. A. (2001).
Linear time inference in hierarchical HMMs.
In *Proceedings of Neural Information Processing Systems*, pages 833–840, Vancouver, BC, Canada.

Bibliography III

Petrides, M. (1987).
Conditional learning and the primate frontal cortex.
In *The frontal lobes revisited*, pages 91–108. IRBN Press, New York.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006).
Learning accurate, compact, and interpretable tree annotation.
In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440. Association for Computational Linguistics.

Roark, B. (2001).
Probabilistic top-down parsing and language modeling.
Computational Linguistics, 27(2):249–276.

Bibliography IV

Schuler, W. (2009).
Parsing with a bounded stack using a model-based right-corner transform.
In *Proceedings of NAACL/HLT 2009*, NAACL '09, pages 344–352, Boulder, Colorado. Association for Computational Linguistics.

Schuler, W., AbdelRahman, S., Miller, T., and Schwartz, L. (2010).
Broad-coverage incremental parsing using human-like memory constraints.
Computational Linguistics, 36(1):1–30.

Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008).
A context-based theory of recency and contiguity in free recall.
Psychological Review, 115:893–912.