A Rational Statistical Parser

Jesús Calvillo and Matthew Crocker

Department of Computational Linguistics and Phonetics
University of Saarland

Outline

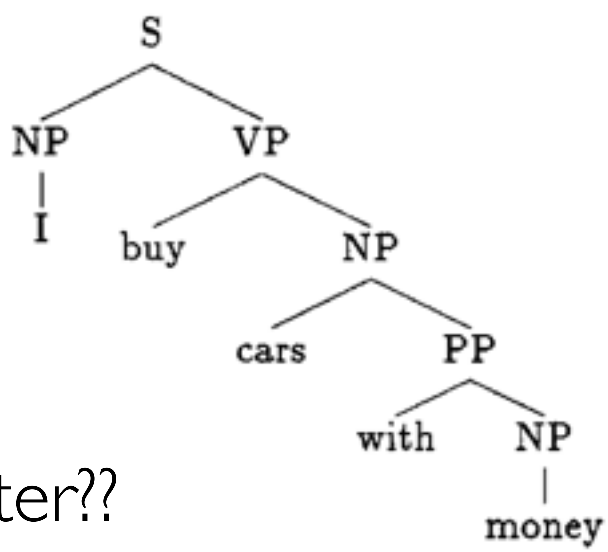
- Introduction
- Rational Parsing Model
- Evaluation
- Conclusion

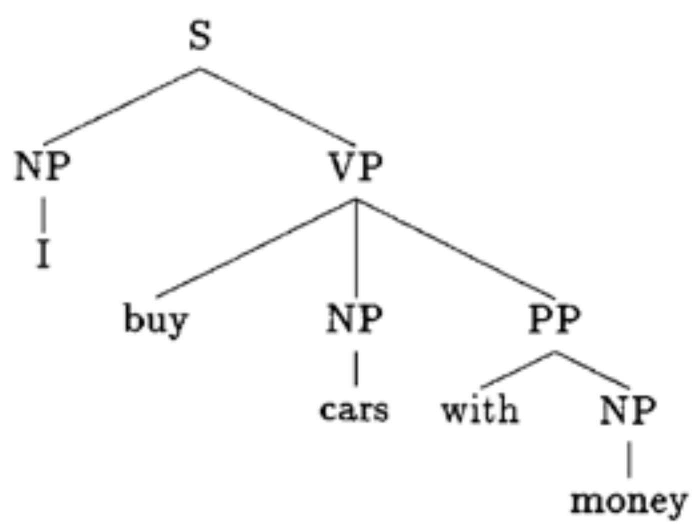
Introduction

When we parse...

sentences → **Syntactic Trees!**

"I buy cars with money."





Which one is better??

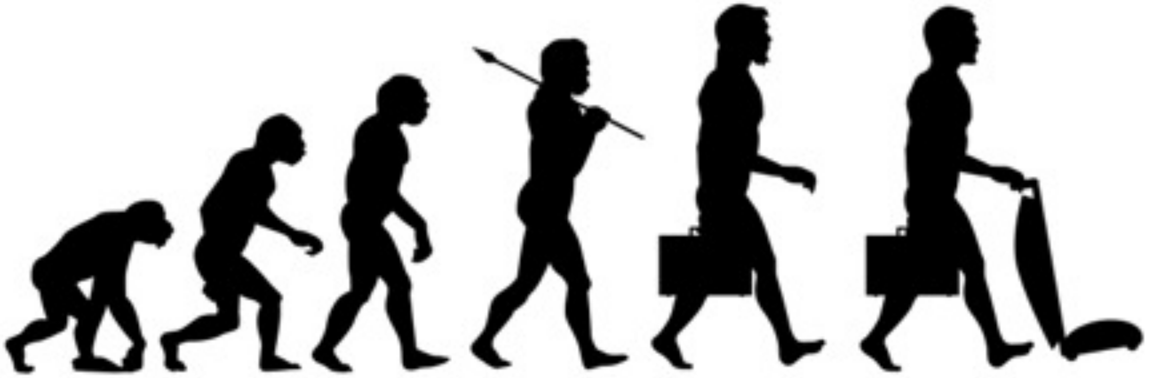
Desiderata

- Trees with higher Probabilities

- Trees with lower Entropies

Principle of Rationality (Anderson, 1991)

The cognitive system optimizes the adaptation of the behavior of the organism.



Anderson(1991)'s Rational Analysis

- **Precisely specify the goals of the system.**
- Develop a model of the environment.
- **Derive the optimal behavioral function given the previous steps.**
- Check if the model is coherent with empirical evidence.

Applications

- **Learning** (Anderson, 1991; Anderson et al., 1997).
- **Vision** (Legge et al., 1997; Liu et al., 1995; Parish and Sperling, 1991; Pelah, 1997).
- **Memory** (Anderson and Milson, 1989; Schooler, 1998; Shanks, 1995).
- **Reasoning** (Cheng, 1997; Oaksford and Chater, 1994)

Rational Syntactic Parsing

Chater et al. (1998)

- Goal:

"To Maximize the probability of recovering the parse of the input as generated by the speaker."

- Balance between 2 risks:
  - Mistakenly rejecting the correct parse
  - "Crashing" during unnecessary search

$$f(H) = P(H) \cdot P(settle\ H) \cdot \frac{1}{1 - P(escape\ H)}$$

Hale (2011)

- Goal:

"Rapidly arriving at the syntactic structure intended by the speaker" as a result of time pressures.

Sentence comprehension should be accurate and quick.

$$\hat{f}(n) = g(n) + \boxed{\hat{h}(n)}$$

→ **A* Search**

Try to achieve good solutions **without wasting too much time** exploring unpromising subspaces.

Hale (2011)

- Number of states visited represent the model's search effort.

- If the heuristic is successful, then fewer states will be explored, otherwise relatively more work is done.

For example,

"The horse raced past the barn fell." → 115 states

"The horse raced past the barn." → 43 states

Hale (2011)

- Similarly, the model predicts:
  - Garden Path Effects (Frazier and Clifton, 1996)
  - Counter examples to the Garden-Path Theory (Gibson, 1991)
  - Local Coherence in English (Tabor et al., 2004) and German (Konieczny, 2005; Konieczny and Müller, 2006)

Parsing Framework

Goal

Retrieve most probable analyses according to experience while saving resources such as time, memory and processing.

Analyses that are shorter or less cognitively complex should be preferred, while maintaining accuracy.

Predicting States

$$S(\omega) = P(\omega) - \beta \cdot EC(\omega)$$

P(w) : information about how probable a derivation is according to past decisions and the current one.

EC(w) : measure of how we expect the rest of the derivation to be.

Hale (2011) – Expected Derivation Length

- As an alternative to average # of steps to goal.

$$expected\ derivation\ length\ (EDL) = A^{\infty} \times \begin{bmatrix} 1 \\ 1 \\ \vdots \end{bmatrix}$$

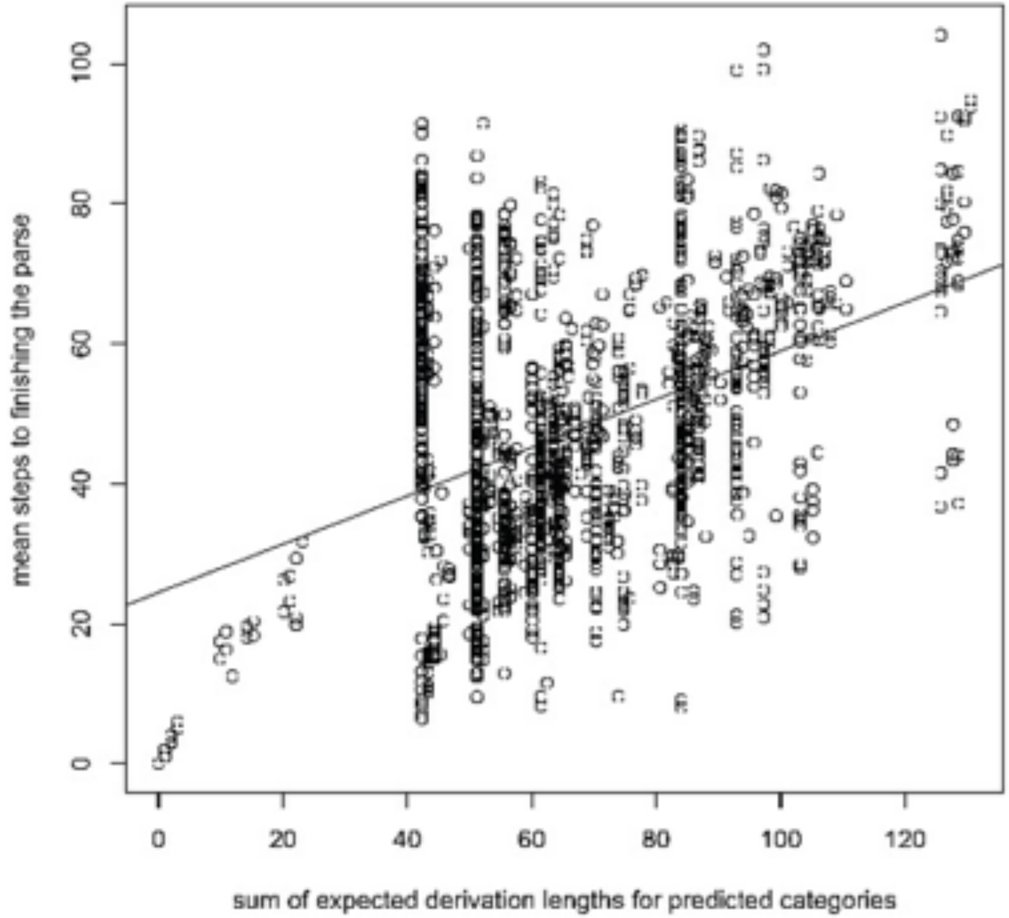
$$A^{\infty} = \sum_{i=0}^{\infty} A^i = \frac{I}{I - A} = (I - A)^{-1}$$

Where *A* is the stochastic expectation matrix.

Each entry *Aij* contains the sum of probabilities that the ith grammar symbol is rewritten as the jth symbol.

Hale (2011) – Expected Derivation Length

- Having each state represented as the sought categories in the stack, the values of $\hat{h}(n)$ correspond **to summing the EDLs of each category in the state *n*.**



Correlation between summed EDLs of Categories sought in a parser and mean number of steps to completion. r=0.4, p<0.0001. Hale(2011).

Hale (2011) – Entropy of a Nonterminal

According to Grenander's theorem (Grenander, 1967), the entropy of a Nonterminal can be estimated:

$$entropy\ of\ a\ nonterminal = A^{\infty} \times \begin{bmatrix} h_S \\ h_{SBAR} \\ h_{NP} \\ h_{NBAR} \\ \vdots \end{bmatrix} \qquad h_x = -\sum_{r \,\epsilon\, R_x} P(r) log_2 P(r)$$

Where *h* is the entropy of a single-rule rewriting event in a derivation of a nonterminal symbol.

The resulting vector corresponds to the average uncertainty values about any derivation started by the given nonterminal.

Entropies, Probabilities and Lengths

- Entropy can also be seen as the negative expectation of a log probability.

$$E_i[log p_r] = \sum_{r \,\epsilon\, R(\phi_i)} p_r log p_r$$

- Applying a similar reasoning, we can say **that the entropy of a nonterminal** corresponds to the **negative expectation of the log probability of a tree whose root is the nonterminal.**

Entropies, Probabilities and Lengths

Expected Derivation Length $\Longleftrightarrow$ Expected Log Probability (-Entropy)

Derivation Length $\Longleftrightarrow$ Log Probability

Models maximizing probabilities and minimizing entropies are at the same time:

✓Minimizing (expected) derivation lengths
✓Maximizing (expected) log probabilities

Surprisal and Entropy Reduction

Entropy Reduction Hypothesis (ERH): The transition from states with high entropies to states with low entropies represents a high cognitive effort. (Hale, 2006).

Similar to ERH, states with high Surprisal are related to high cognitive effort (Hale, 2001).

Both are related to the disambiguation effort the parser performs in view of new information.

Surprisal and Entropy Reduction

$$Surprisal(\tau) = -logP(\tau)$$

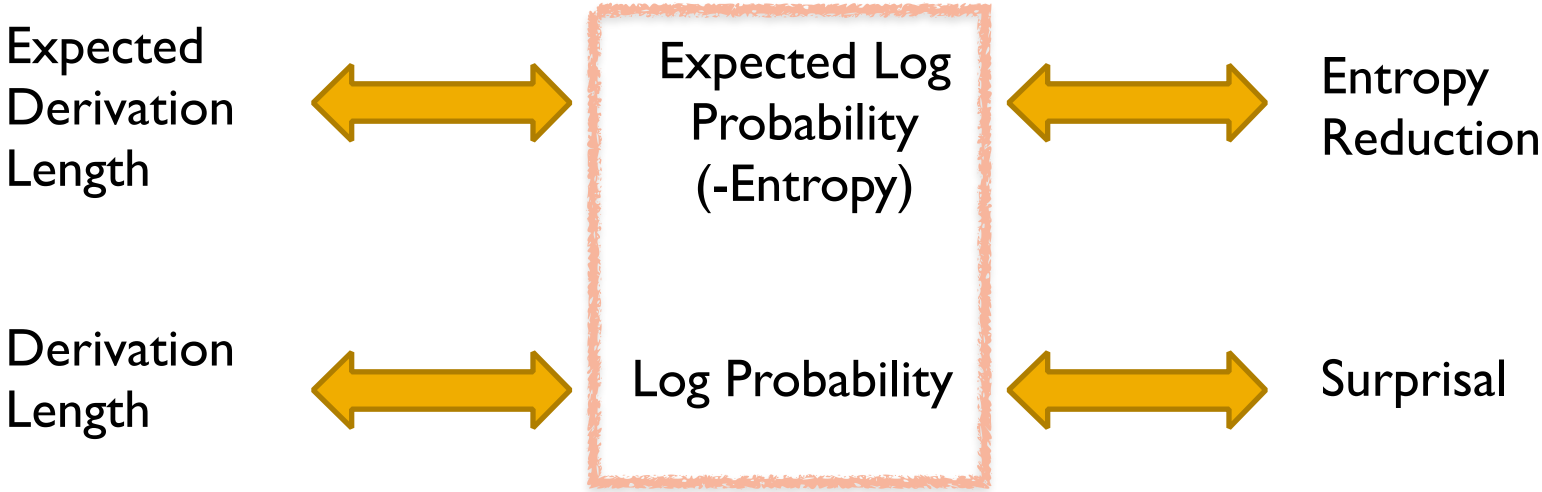
The definition of entropy is equal to the expected value of Surprisal.

Hence, Roark (2011) dubbed it Expected Surprisal:

$$ExpectedSurprisal(\tau_{NT}) = -E[logP(\tau_{NT})] = Entropy(NT)$$

$$\hat{\tau} = \arg\max_{\tau \epsilon T}(logP(\tau) + E[logP(\tau)]) = \arg\min_{\tau \epsilon T}(Surprisal(\tau) + Entropy(\tau))$$

Goals



Taking these points of view, the models minimizing entropies and maximizing probabilities achieve the following goals:

✓Being quick
✓**Retrieving the most probable analyses according to experience**
✓Retrieving cognitively "easy" analyses → saving resources

Cognitive Load Minimization

Parsing Model

Cognitive Load Minimization

$$S(\tau) = logP(\tau) \; - \; \beta \cdot Ent(\tau)$$

- Trade-off between 2 kinds of Cognitive Load:
  - Surprisal
  - Entropy Reduction

The resulting score corresponds to a minimization of both measures such that the performance in the form of f-scores is improved.

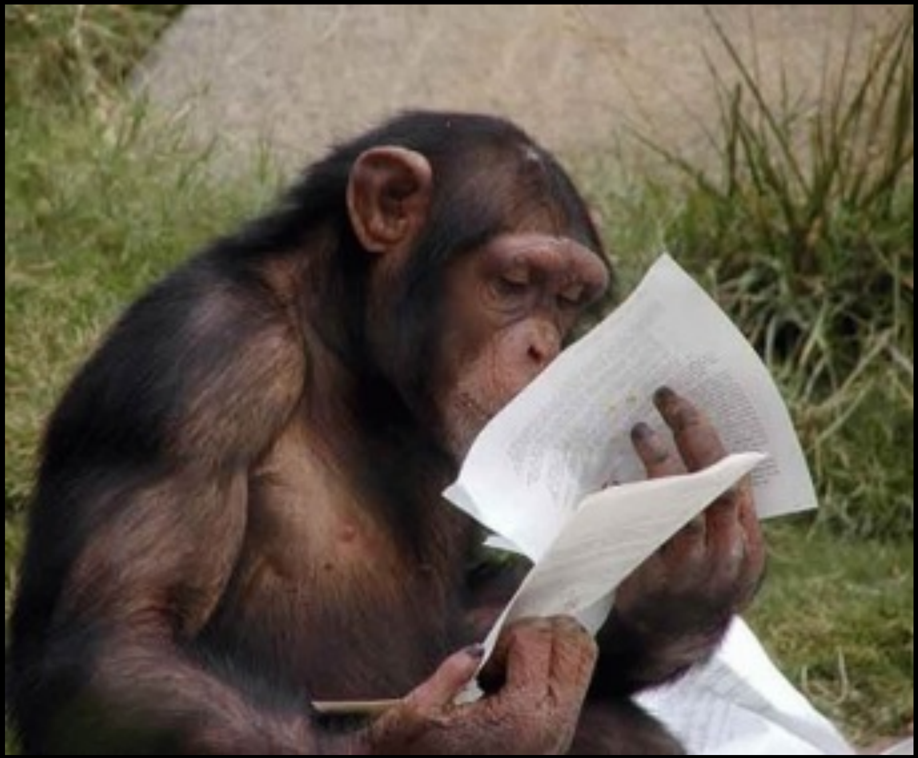
Cognitive Load Minimization – Assumptions

- During production, speakers are cognitively constrained and prone to generate cognitively manageable constructions.
- To maximize the probability of successful communication, speakers should be able to modulate the complexity of their utterances such that the comprehender is able to follow.
- **Normal/understandable sentences should show a bias towards cognitively manageable constructions.**

Implementation

- Augmented Version of the system by Roger Levy (2008).
- Implementation of the Stolcke Parser (Stolcke, 1995).
- Extension of the Earley Parsing Algorithm (Earley, 1970).
- Viterbi Retrieval of N-Best Parses.

Evaluation



Cognitive Load Minimizing Model

- Model that minimizes **Syntactic Surprisal** and **Entropy**

$$S(\tau) = \sum_{r \,\epsilon\, R(\tau)} logP(r) - \beta \cdot \sum_{nt \,\epsilon\, NT_\tau} H(nt)$$

$$\hat{\tau} = \arg\max_{\tau \,\epsilon\, T} S(\tau)$$

Grammar

- The grammar was extracted from sections 02-22 of the Penn Treebank (Marcus et al., 1993).
- No markovization, lexicalization nor parent annotation.

Finding *H(nt)*

- Development Set:
  Section 24 of the Penn Treebank

1) Retrieve the N-best parses of each sentence according only to probabilities.

2) Re-rank them according to different variations of the model.

3) Final formulation is the one that provides maximum benefit in terms of f-score.

$H(nt) \rightarrow$ Localized Entropy

- Contribution of each nonterminal is in function of its location within the tree.
- Distance between the location of the NT and the root of the tree according to a preorder representation.

For example:

$$(ROOT(S(NP(DT)(NNS))(ADVP(RB))(VP(VBD))(.)))$$

Distance(NP) = 1 (there is 1 nonterminal between NP and ROOT).

Distance(ADVP)=4

$H(nt) \rightarrow$ *Localized Entropy*

- *Localized entropy* is the score that each nonterminal NT will add to the final entropy score of the tree:

$$h_L = \frac{h(NT)}{Distance(root, NT)^{\alpha}}$$

where *h(NT)* is taken directly from the grammar and **α** is another parameter. In practice we found 2 to be a good exponent.

- Having defined h_L, the entropy score *H(T)* of a tree *T* is the following:

$$H(T) = \sum_{NT \epsilon T} h_L(NT)$$

But we don't have the preorder representation of the tree!!

Preorder Approximation

$$Distance(root, nt) = 2i + (|\, x \,| - j)$$

where
(i,j) : coordinates of *nt*
|x|: size of the sentence.

Nodes closer to the beginning of the sentence have less distance.

Nodes closer to the root of the tree have less distance.

		a	circle	touches	a	square
0	$_0 \rightarrow .S$ *predicted* $_0$S→.NP VP $_0$NP→ .Det N $_0$Det→ .a					
1	*scanned* $_0$Det→ a. *completed* $_0$NP→Det. N	*predicted* $_1$N→ .circle $_1$N→ .square $_1$N→ .triangle				
2	*completed* $_0$NP→ Det N. $_0$S→NP. VP	*scanned* $_1$N→ circle.	*predicted* $_2$VP→ .VT NP $_2$VP→ .VI PP $_2$VT→ .touches $_2$VI→ .is			
3			*scanned* $_2$VT→ touches. *completed* $_2$VP→VT. NP	*predicted* $_3$NP→ .Det N $_3$Det→.a		
4				*scanned* $_3$Det→ a. *completed* $_3$NP→Det. N	*predicted* $_4$N→ .circle $_4$N→.square $_4$N→.triangle	
5	*completed* $_0$S→NP VP. $_0 \rightarrow$S.		*completed* $_2$VP→VT NP.	*completed* $_3$NP→Det N.	*scanned* $_4$N→ square.	
	0	1	2	3	4	5

Final Formulation

$$Distance(root, nt) = 2i + (\mid x \mid - j)$$

$$h_L = \frac{h(NT)}{Distance(root, NT)^{\alpha}}$$

$$S(\tau) = \sum_{r \, \epsilon \, R(\tau)} logP(r) - 0.6 \cdot \sum_{nt \, \epsilon \, NT_{\tau}} h_L(nt)$$

About the Final Formulation

- As the parser goes further on the sentence and deeper on the tree, the relevance of Entropy decreases.
- At the beginning of production speakers construct the syntactic structure to be as concise/probable as possible.
- As the speaker continues, the possible continuations of each prefix become less and less as semantic and syntactic restrictions arise.
- As the parser advances, the need to be succinct should decrease, as the sentence is about to reach the end anyways.

About the Final Formulation

- Even though the relevance of entropy decreases, it never reaches zero.

- For structures with equal probabilities, the parser would still prefer the one with lowest entropy.

Final Experiment

- Test Set: Section 23 of the Penn Treebank (2416 sentences).
- Task: Obtain the best parse for each sentence using our model against using only traditional PCFG probabilities.

Results - First Analysis

	Baseline	P-H	Δ
All Sentences			
Number of Sentences	2416	2416	
Bracketing Recall	70.39	**71.41**	+1.02
Bracketing Precision	75.73	**76.94**	+1.21
Bracketing F-Score	72.96	**74.07**	+1.11
Complete Match	9.73	10.10	+0.37
Average Crossing	2.89	2.67	-0.22
No Crossing	33.57	35.64	+2.07
2 or less Crossing	59.27	63.16	+3.89
Length $\leq$ *40*			
Number of Sentence	2245	2245	
Bracketing Recall	71.86	**72.86**	+1.0
Bracketing Precision	77.13	**78.33**	+1.2
Bracketing F-Score	74.40	**75.49**	+1.09
Complete Match	10.47	10.87	+0.4
Average Crossing	2.47	2.27	-0.2
No Crossing	35.99	38.17	+2.18
2 or less Crossing	62.85	66.95	+4.1

Results - Second Analysis

- Output of the parser converted to dependency trees using the tool by Johansson and Nugues (2007).
- Output was evaluated using the CoNLL-07 shared task evaluation script.

	Baseline	P-H	Δ
Labeled attachment score:	75.29	**76.13**	+0.84
Unlabeled attachment score:	77.80	**78.66**	+0.86
Label accuracy score:	82.64	**83.08**	+0.44

Results – Dependency Relations

		Baseline				Probability - Entropy (P-H)				Δ			
Dep.	gold	correct	system	recall	precision	correct	system	recall	precision	correct	system	recall	precision
ADV	4085	3269	5183	80.02	63.07	3235	5154	79.19	62.77	34	29	0.83	0.30
AMOD	980	536	979	54.69	54.75	529	974	53.98	54.31	7	5	0.71	0.44
CC	188	172	190	91.49	90.53	172	190	91.49	90.53	0	0	0.00	0.00
COORD	2795	2149	2824	76.89	76.10	2125	2821	76.03	75.33	24	3	0.86	0.77
DEP	1072	916	1215	85.45	75.39	917	1216	85.54	75.41	-1	-1	-0.09	-0.02
IOBJ	296	99	377	33.45	26.26	100	375	33.78	26.67	-1	2	-0.33	-0.41
NMOD	19515	16423	17707	84.16	92.75	16347	17724	83.77	92.23	76	-17	0.39	0.52
OBJ	3497	2106	3229	60.22	65.22	2101	3235	60.08	64.95	5	-6	0.14	0.27
P	6870	6844	6876	99.62	99.53	6844	6876	99.62	99.53	0	0	0.00	0.00
PMOD	5574	4577	5537	82.11	82.66	4522	5534	81.13	81.71	55	3	0.98	0.95
PRD	671	526	664	78.39	79.22	527	666	78.54	79.13	-1	-2	-0.15	0.09
PRN	140	69	115	49.29	60.00	69	116	49.29	59.48	0	-1	0.00	0.52
PRT	159	159	180	100.00	88.33	159	180	100.00	88.33	0	0	0.00	0.00
ROOT	2416	2010	2416	83.20	83.20	1978	2416	81.87	81.87	32	0	1.33	1.33
VC	1871	1700	2282	90.86	74.50	1708	2276	91.29	75.04	-8	6	-0.43	-0.54
VMOD	6555	5540	6910	84.52	80.17	5508	6931	84.03	79.47	32	-21	0.49	0.70

- Most frequent dependencies are treated better by the new model.
- Infrequent dependencies are treated equally by both models.
- Slightly frequent relations present a worse performance with the new model.

Conclusion

- We provided a definition of syntactic parsing at the computational level as a trade-off between probability and entropy.

- The function that we utilized has some peculiarities, namely, the weight of entropies decreases as the parser goes along on the sentence and deeper in the syntactic tree.

Conclusion

- We used the notions of surprisal and entropy reduction as cognitive load measures.
- We assumed that syntactic analyses with low cognitive load should be ranked higher.
- The resulting system showed a modest but general improvement over the baseline.

Future Work

- Systematic and extensive trials with other functions to combine entropies.
- Trials with more complex grammars.
- Experiments with sentences presenting psycholinguistic phenomena such as Garden Path, Local Coherence, etc.
- Experiments with other languages.

Thanx !!!

References

- John Anderson. The place of cognitive architectures in rational analysis. In K. VanLehn, editor, Architectures for Cognition. Lawrence Erlbaum, Hillsdale, NJ, 1991.
- John R Anderson and Robert Milson. Human memory: An adaptive perspective. Psychological Review, 96(4): 703, 1989.
- Richard B Anderson, Ryan D Tweney, Mark Rivardo, and Sean Duncan. Need probability affects retention: A direct demonstration. Memory & cognition, 25(6):867-872, 1997.
- Nicholas Chater, Matthew J Crocker, and Martin J Pickering. The rational analysis of inquiry: The case of parsing. 1998.
- Patricia W Cheng. From covariation to causation: A causal power theory. Psychological review, 104(2):367, 1997.
- Jay Earley. An ecient context-free parsing algorithm. Communications of the ACM, 13(2):94{102, 1970.
- John Hale. Uncertainty about the rest of the sentence. Cognitive Science, 30(4):643{672, 2006.
- John T. Hale. What a rational parser would do. Cognitive Science, 35(3):399{443, 2011.
- Richard Johansson and Pierre Nugues. Extended constituent-to-dependency conversion for English. In Proceedings of NODALIDA 2007, pages 105{112, Tartu, Estonia, May 25-26 2007.
- Gordon E Legge, Timothy S Klitz, and Bosco S Tjan. Mr. chips: an ideal-observer model of reading. Psychological review, 104(3):524, 1997.
- Roger Levy. Expectation-based syntactic comprehension. Cognition, 106(3):1126{1177, 2008.

References

- Zili Liu, David C Knill, and Daniel Kersten. Object classication for human and ideal observers. Vision research, 35(4):549-568, 1995.
- Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. Computational linguistics, 19(2): 313-330, 1993.
- David Marr. Vision. Freeman, New York, 1982.
- Mike Oaksford and Nick Chater. A rational analysis of the selection task as optimal data selection. Psychological Review, 101(4):608, 1994.
- David H Parish and George Sperling. Object spatial frequencies, retinal spatial frequencies, noise, and the eciency of letter discrimination. Vision research, 31(7): 1399-1415, 1991.
- Adar Pelah. The vision of natural and complex images. Vision research, 37(23):3201-3202, 1997.
- L Schooler. Sorting out core memory processes. Rational Models of Cognition, Oxford University Press, Oxford, pages 128-55, 1998.
- David R Shanks. Is human learning rational? The Quarterly Journal of Experimental Psychology, 48(2): 257-279, 1995.
- Andreas Stolcke. An ecient probabilistic context-free parsing algorithm that computes prefix probabilities. Computational linguistics, 21(2):165-201, 1995.
- David Vadas. Statistical parsing of noun phrase structure. 2010.

Entropy of Nonterminals

According to Grenander's Theorem:

$$h_i = h(\phi_i) = - \sum_{r \, \epsilon \, R(\phi_i)} p_r log_2 p_r$$

$$H(\phi_i) = h(\phi_i) + \sum_{r \, \epsilon \, R(\phi_i)} p_r [H(\phi_{j1}) + H(\phi_{j2}) + ...]$$

First term is the definition of entropy for a random variable.

Second term is the recurrence. It expresses the intuition that derivational uncertainty is propagated from children to parents (Hale, 2006).