Baby Steps: How "Less is More" in Unsupervised Dependency Parsing

VALENTIN I. SPITKOVSKY, HIYAN ALSHAWI AND DANIEL JURAFSKY

vals@stanford.edu, hiyan@google.com and jurafsky@stanford.edu

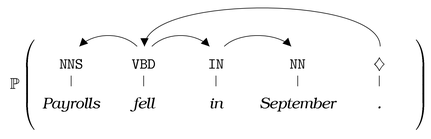
SUMMARY

We present two simple, data complexity aware approaches for unsupervised grammar induction and apply them to Klein and Manning's Dependency Model with Valence: **Baby Steps** bootstraps itself without a prior, by increasing data complexity gradually; **Less is More** focuses on fewer but lower complexity examples, trading off quantity against ambiguity.

DEPENDENCY MODEL WITH VALENCE

- generative process
- projective trees
- single-state head automata
- unlexicalized (part-of-speech tags)

EXAMPLE:



$$\begin{aligned}
= \; & (1 - \mathbb{P}_{\text{STOP}}(\diamond, \text{L}, \text{T}))^{0} \\
\times \; & (1 - \mathbb{P}_{\text{STOP}}(\text{VBD}, \text{R}, \text{T})) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{VBD}, \text{L}, \text{F}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{IN}, \text{L}, \text{T}) \\
\times \; & \mathbb{P}_{\text{ATTACH}}(\diamond, \text{L}, \text{VBD}) \\
\times \; & \mathbb{P}_{\text{ATTACH}}(\text{VBD}, \text{R}, \text{IN}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{VBD}, \text{R}, \text{F}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{IN}, \text{R}, \text{F}) \\
\times \; & (1 - \mathbb{P}_{\text{STOP}}(\text{VBD}, \text{L}, \text{T})) \\
\times \; & (1 - \mathbb{P}_{\text{STOP}}(\text{IN}, \text{R}, \text{T})) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{NNS}, \text{L}, \text{T}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{NN}, \text{L}, \text{T}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\diamond, \text{L}, \text{F})^{1} \\
\times \; & \mathbb{P}_{\text{ATTACH}}(\text{VBD}, \text{L}, \text{NNS}) \\
\times \; & \mathbb{P}_{\text{ATTACH}}(\text{IN}, \text{R}, \text{NN}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{NNS}, \text{R}, \text{T}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\text{NN}, \text{R}, \text{T}) \\
\times \; & \mathbb{P}_{\text{STOP}}(\diamond, \text{R}, \text{T})^{1}.
\end{aligned}$$

STANDARD TRAINING: **Ad-Hoc** Init ⇒ WSJ10

- EM via inside-outside re-estimation (on WSJ10)
- Ad-Hoc harmonic initializer aims for balanced trees

CONCLUSION

Small tweaks to Klein and Manning's approach of 2004 beat 2009 state-of-the-art on longer sentences. We suspect that other parsing models and unsupervised learning algorithms also stand to gain from an awareness of data complexity.

ACKNOWLEDGMENTS

The National Science Foundation, award IIS-0811974.

The Fannie & John Hertz Foundation, for unwavering, generous support and encouragement.

Prof. Christopher D. Manning, for a wonderful course on "Natural Language Processing" (CS.224N), and **Prof. Andrew Y. Ng**, for an instantly useful and accessible course on "Machine Learning" (CS.229).

BABY STEPS

No Init ⇒ WSJ1 → WSJ2 → ⋯ → WSJ45

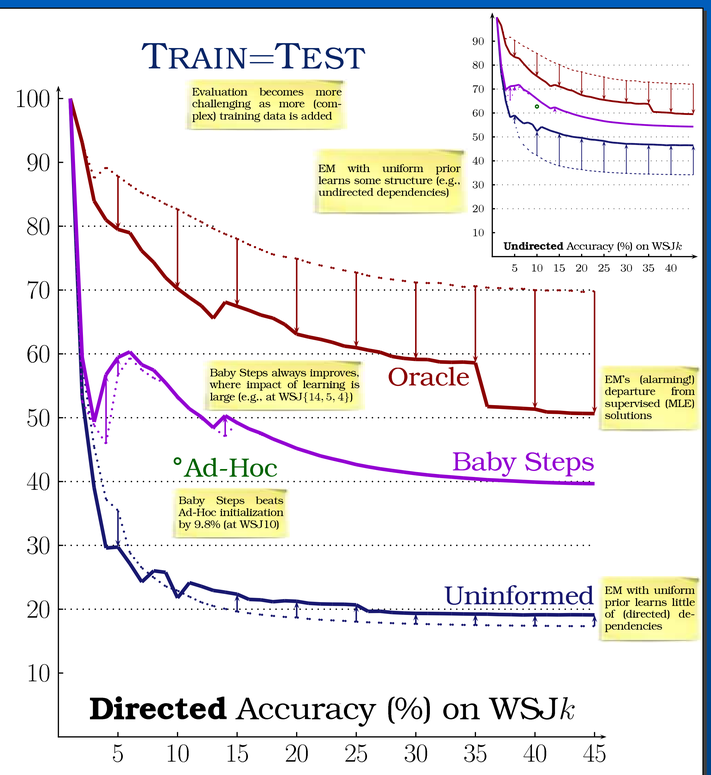
- iterative training with increasingly longer sentences
- initialization-free

LESS IS MORE

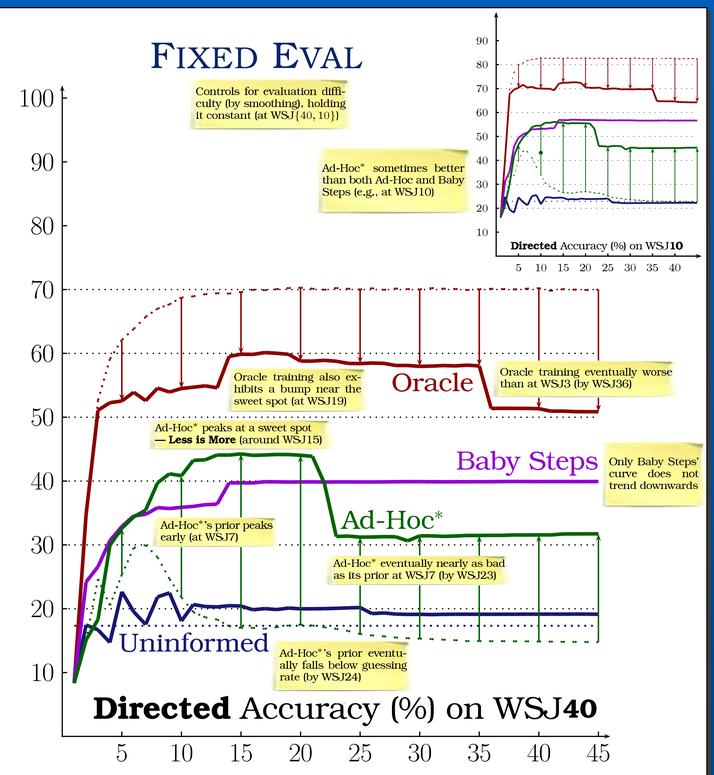
Ad-Hoc* Init ⇒ WSJ15

- standard batch training, but at "sweet spot" lengths
- linguistically-informed prior (our implementation of Ad-Hoc)

TRAIN=TEST



FIXED EVAL



GENERALIZATION

	*Ad-Hoc**	*Baby Steps*		*Ad-Hoc**	*Baby Steps*	
Section 23	**44.1** (58.8)	39.2 (53.8)	@15	31.5 (51.6)	39.4 (54.0)	@45
WSJ100	**43.8** (58.6)	39.2 (53.8)		31.3 (51.5)	39.4 (54.1)	
Brown100	**43.3** (59.2)	42.3 (55.1)		32.0 (52.4)	42.5 (55.5)	

NIPS Workshop on Grammar Induction, Representation of Language and Language Learning

Neural Information Processing Systems Foundation

December 11, 2009

Whistler, Canada

VENUE

STATE OF THE ART

			Decoding	WSJ10	WSJ20	WSJ∞
	Attach-Right	(Klein and Manning, 2004)	—	38.4	33.4	31.7
DMV	Ad-Hoc	(Klein and Manning, 2004)	Viterbi	45.8	39.1	34.2
	Dirichlet	(Cohen et al., 2008)	Viterbi	45.9	39.4	34.9
	Ad-Hoc	(Cohen et al., 2008)	MBR	46.1	39.9	35.9
	Dirichlet	(Cohen et al., 2008)	MBR	46.1	40.6	36.9
	Log-Normal Families	(Cohen et al., 2008)	Viterbi	59.3	45.1	39.0
	Baby Steps (@15)		*Viterbi*	*55.5*	*44.3*	*39.2*
	Baby Steps (@45)		*Viterbi*	*55.1*	*44.4*	*39.4*
	Log-Normal Families	(Cohen et al., 2008)	MBR	59.4	45.9	40.5
	Shared Log-Normals (tie-verb-noun)	(Cohen and Smith, 2009)	MBR	61.3	47.4	41.4
	Bilingual Log-Normals (tie-verb-noun)	(Cohen and Smith, 2009)	MBR	62.0	48.0	42.2
	Less is More (Ad-Hoc* @15)		*Viterbi*	*56.2*	**48.2**	**44.1**
EVG	Smoothed (skip-val)	(Headden et al., 2009)	Viterbi	62.1		
	Smoothed (skip-head)	(Headden et al., 2009)	Viterbi	65.0	*Section 23*	
	Smoothed (skip-head), Lexicalized	(Headden et al., 2009)	Viterbi	**68.8**		