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

EXAMPLE:

\[
\begin{align*}
\text{p} & = \text{Rainy}(\text{p}, \text{lastyear}) \\
\text{p} & = \text{Hot}(\text{p}, \text{September})
\end{align*}
\]

STANDARD TRAINING: Ad-Hoc Int - WSJ10

- EM via inside-outside re-estimation (on WSJ10)
- Ad-Hoc harmonic initialization 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 authors acknowledge support from Google Inc. and Stanford University. Prof. Christopher D. Manning is currently serving on “Natural Language Processing” (NLP) 2012, and Prof. Andrew W. N. for an invitation to attend MINT (Multilingual Information Retrieval).