Three Dependency-and-Boundary Models for Grammar Induction

Valentin I. Spitkovsky, Hiyan Alshawi and Daniel Jurafsky

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

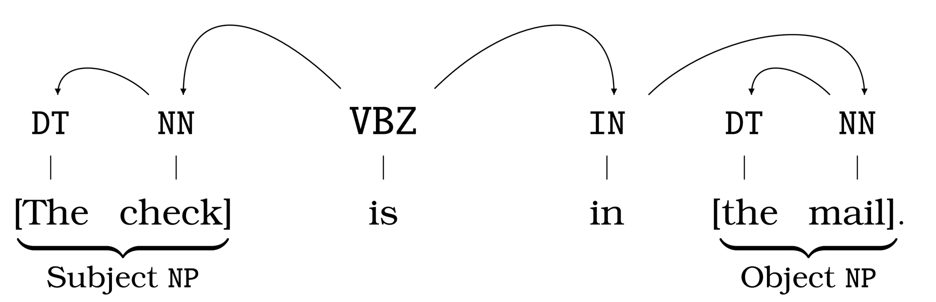
Idea

Use boundary cues in head-driven dependency grammars.

Intuition

Induce structure by working inwards from edges.

Example:



- learn from left fringe (determiner DT) how to parse object NP
- based on right fringe (noun NN), correctly parse subject NP
- between the two, glean make-up of larger phrases (e.g., VP)

Motivation

Machine Learning: Focus on Observable, Complementary Features

- weak equivalence of phrase representations (Xia and Palmer, 2001)
- redundant views of data ease learning (Blum and Mitchell, 1998)

Grammar Induction: A Bounty of Constituent Boundary Markers

- at sentence beginnings and ends (Hänig et al., 2008; Hänig, 2010)
- around function words (Berant et al., 2006)
- around punctuation marks (Seginer, 2007) (Ponvert et al., 2010; 2011) (Spitkovsky et al., 2011)
- at capitalization (or script) change-points (Spitkovsky et al., 2012)
- at web markup bracketing end-points (Spitkovsky et al., 2010)
- around other semantic annotations (Naseem and Barzilay, 2011)

Language Acquisition: Word Boundaries Matter to Human Learning

- importance of exposure to isolated words (Brent and Siskind, 2001)

DBM-1

Stop generation based on fringe words of partial yields.

$$\mathbb{P}_{\texttt{STOP}}(\,\cdot \mid dir;\ adj, c_e)$$

Example: *(The check is in the mail.)*

$$\begin{aligned}
\mathbb{P} = {} & (1 - \overbrace{\mathbb{P}_{\texttt{STOP}}(\diamond \mid \texttt{L};\ \texttt{T})}^{0}) & & \times \mathbb{P}_{\texttt{ATTACH}}(\texttt{VBZ} \mid \diamond;\ \texttt{L}) \\
& \times (1 - \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{T}, \texttt{VBZ})) & & \times \mathbb{P}_{\texttt{ATTACH}}(\texttt{NN} \mid \texttt{VBZ};\ \texttt{L}) \\
& \times (1 - \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{T}, \texttt{VBZ})) & & \times \mathbb{P}_{\texttt{ATTACH}}(\texttt{IN} \mid \texttt{VBZ};\ \texttt{R}) \\
& \times \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{F}, \texttt{DT}) \quad \texttt{// VBZ} & & \times \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{F}, \texttt{NN}) \quad \texttt{// VBZ} \\
& \times (1 - \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{T}, \texttt{NN}))^2 & & \times \mathbb{P}^2_{\texttt{ATTACH}}(\texttt{DT} \mid \texttt{NN};\ \texttt{L}) \\
& \times (1 - \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{T}, \texttt{IN})) & & \times \mathbb{P}_{\texttt{ATTACH}}(\texttt{NN} \mid \texttt{IN};\ \texttt{R}) \\
& \times \mathbb{P}^2_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{T}, \texttt{NN}) & & \times \mathbb{P}^2_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{F}, \texttt{DT}) \quad \texttt{// NN} \\
& \times \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{T}, \texttt{IN}) & & \times \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{F}, \texttt{NN}) \quad \texttt{// IN} \\
& \times \mathbb{P}^2_{\texttt{STOP}}(\,\cdot \mid \texttt{L};\ \texttt{T}, \texttt{DT}) & & \times \mathbb{P}^2_{\texttt{STOP}}(\,\cdot \mid \texttt{R};\ \texttt{T}, \texttt{DT}) \\
& \times \underbrace{\mathbb{P}_{\texttt{STOP}}(\diamond \mid \texttt{L};\ \texttt{F})}_{1} & & \times \underbrace{\mathbb{P}_{\texttt{STOP}}(\diamond \mid \texttt{R};\ \texttt{T})}_{1}.
\end{aligned}$$

- truly head-outward model (Alshawi, 1996)
- still split-head, hence efficient (Eisner and Satta, 1999)
- conditions on more observable state — left and right words of phrases being constructed — than hidden head words

→ well-suited to unsupervised learning

DBM-2

Models incomplete inputs based on boundary punctuation.

$$\mathbb{P}_{\texttt{ATTACH}}(c_r \mid \diamond;\ \texttt{L}, comp) \quad \text{and} \quad \mathbb{P}_{\texttt{STOP}}(\,\cdot \mid dir;\ adj, c_e, comp)$$

Examples: *(Ungrammatical news-style fragments.)*

Odds and Ends	captions and headlines
George Morton	proper noun phrases
Revenue: $3.57 billion	monetary values
c - Domestic car	line items
1:11am	date and time expressions

- incomplete fragments are uncharacteristically short
- roots of fragments are generally not verbs or modals
- have multiple overlapping grammars coexist in a model

→ avoid pitfalls, like inducing nouns as sentence heads

DBM-3

Incorporates sentence-internal punctuation boundaries.

$$\mathbb{P}_{\texttt{ATTACH}}(c_d \mid c_h;\ dir, cross)$$

Example: *Continentals believe that the strongest growth area will be southern Europe.*

- punctuation-crossing vets common remote constructions
- e.g., subordinating conjunctions (IN) and their dependent modal verbs (MD), which are, on average, 4.8 tokens apart
- avoids bad long distance relations (e.g., far-off DT-NN pairs)

→ learn to piece together inter-punctuation fragments

Summary

Unsupervised split-head dependency grammars:

GB (Paskin, 2001); DMV (Klein and Manning, 2004); EVG (Headden et al., 2009).

	$\mathbb{P}_{\texttt{ATTACH}}$	$\mathbb{P}_{\texttt{ATTACH}}$	$\mathbb{P}_{\texttt{STOP}}$
GB	$1 / \lvert\{w\}\rvert$	$d \mid h;\ dir$	$1 / 2$
DMV	$c_r \mid \diamond;\ \texttt{L}$	$c_d \mid c_h;\ dir$	$\cdot \mid dir;\ adj,\ c_h$
EVG	$c_r \mid \diamond;\ \texttt{L}$	$c_d \mid c_h;\ dir,\ adj$	$\cdot \mid dir;\ adj,\ c_h$
DBM-1	$c_r \mid \diamond;\ \texttt{L}$	$c_d \mid c_h;\ dir$	$\cdot \mid dir;\ adj,\ c_e$
DBM-2	$c_r \mid \diamond;\ \texttt{L},\ comp$	$c_d \mid c_h;\ dir$	$\cdot \mid dir;\ adj,\ c_e,\ comp$
DBM-3	$c_r \mid \diamond;\ \texttt{L},\ comp$	$c_d \mid c_h;\ dir,\ cross$	$\cdot \mid dir;\ adj,\ c_e,\ comp$
	(head-root)	*(dependent-head)*	*(direction, adjacency)*

Results

Previous state-of-the-art: 38.2% directed dependency accuracy.

— average over all 19 languages of the 2006/7 CoNLL sets (Spitkovsky et al., 2011)

- DBM-1: 40.7% (uniform initialization, no predefined input length cutoff)
- DBM-1 → DBM-2 → DBM-3: **42.9**% (staged curriculum training)

Acknowledgments

Partially funded by Defense Advanced Research Projects Agency (DARPA) Machine Reading Program, under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. We thank Roi Reichart and Marta Recasens — as well as Marie-Catherine de Marneffe, Roy Schwartz and Mengqiu Wang — for helpful comments on draft versions of our paper.