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

Valentin I. Spitkovsky

with **Hiyan Alshawi** (Google Inc.)
and **Daniel Jurafsky** (Stanford University)

Idea: (At Least) Two Axes worth Scaffolding

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity**

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity [classic NLP]**

  — word alignment (unsupervised), e.g., IBM models 1-5
  (Brown et al., 1993)

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity** [classic NLP]

  — word alignment (unsupervised), e.g., IBM models 1-5
  (Brown et al., 1993)

  — parsing (supervised), e.g., "coarse-to-fine" grammars
  (Charniak and Johnson, 2005; Petrov 2009)

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity [classic NLP]**

  — word alignment (unsupervised), e.g., IBM models 1-5
  (Brown et al., 1993)

  — parsing (supervised), e.g., "coarse-to-fine" grammars
  (Charniak and Johnson, 2005; Petrov 2009)

- **Data (or Problem / Task) Complexity**

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity [classic NLP]**

  — word alignment (unsupervised), e.g., IBM models 1-5
  (Brown et al., 1993)

  — parsing (supervised), e.g., "coarse-to-fine" grammars
  (Charniak and Johnson, 2005; Petrov 2009)

- **Data (or Problem / Task) Complexity [rare in NLP]**

  — reinforcement learning, e.g., robot navigation
  (Singh, 1992; Sanger 1994)

Idea: (At Least) Two Axes worth Scaffolding

- **Model (or Algorithmic) Complexity [classic NLP]**

  — **word alignment (unsupervised), e.g., IBM models 1-5** (Brown et al., 1993)

  — **parsing (supervised), e.g., "coarse-to-fine" grammars** (Charniak and Johnson, 2005; Petrov 2009)

- **Data (or Problem / Task) Complexity [rare in NLP]**

  — **reinforcement learning, e.g., robot navigation** (Singh, 1992; Sanger 1994)

  — **closest in NLP: cautious named entity classification** (Collins and Singer, 1999; Yarowsky, 1995)

Outline: Three Data-Complexity-Aware Techniques

Outline: Three Data-Complexity-Aware Techniques

- **Baby Steps: scaffolding on data complexity — iterative, requires no initialization**

Outline: Three Data-Complexity-Aware Techniques

- **Baby Steps: scaffolding on data complexity**
  — iterative, requires no initialization

- **Less is More: filtering by data complexity**
  — batch, capable of using a good initializer

Outline: Three Data-Complexity-Aware Techniques

- **Baby Steps: scaffolding on data complexity**
  — iterative, requires no initialization

- **Less is More: filtering by data complexity**
  — batch, capable of using a good initializer

- **Leapfrog: a combination (best of both worlds)**
  — intended as an efficiency hack (but performs best)

Problem: Unsupervised Learning of Parsing

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

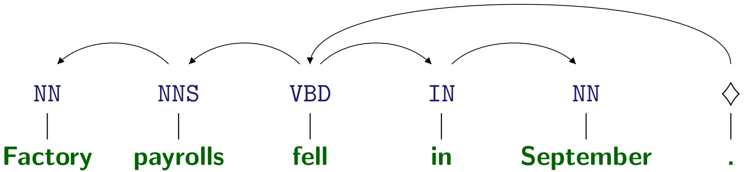
NN	NNS	VBD	IN	NN	♢
Factory	**payrolls**	**fell**	**in**	**September**	**.**

Problem: Unsupervised Learning of Parsing

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

- **Output: Syntactic Structures (and a Probabilistic Grammar)**



Motivation: Unsupervised (Dependency) Parsing

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks:**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**

- **... adjust for any data without reference tree banks: — i.e., exotic languages**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**

- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**

- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;
  - **web search**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;
  - **web search**
    — retrieval, query refinement;

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;
  - **web search**
    — retrieval, query refinement;
  - **question answering**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., exotic languages and/or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;
  - **web search**
    — retrieval, query refinement;
  - **question answering, speech recognition, etc.**

State-of-the-Art: Directed Dependency Accuracy

State-of-the-Art: Directed Dependency Accuracy

42.2% on Section 23 (all sentences) of WSJ

(Cohen and Smith, 2009)

State-of-the-Art: Directed Dependency Accuracy

42.2% on Section 23 (all sentences) of WSJ

(Cohen and Smith, 2009)

31.7% for the (right-branching) baseline

(Klein and Manning, 2004)

State-of-the-Art: Directed Dependency Accuracy

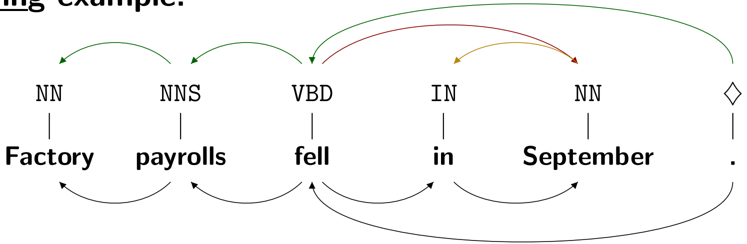
42.2% on Section 23 (all sentences) of WSJ

(Cohen and Smith, 2009)

31.7% for the (right-branching) baseline

(Klein and Manning, 2004)

Scoring example:



Directed Score: $\frac{3}{5} = 60\%$ (**baseline**: $\frac{2}{5} = 40\%$);

Undirected Score: $\frac{4}{5} = 80\%$ (**baseline**: $\frac{4}{5} = 80\%$).

State-of-the-Art: A Brief History

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

- **2004 — annealing techniques** **(Smith and Eisner)**

State-of-the-Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)
- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

- 2004 — annealing techniques (Smith and Eisner)
- 2005 — contrastive estimation (Smith and Eisner)

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

- **2004 — annealing techniques** **(Smith and Eisner)**
- **2005 — contrastive estimation** **(Smith and Eisner)**
- **2006 — structural biasing** **(Smith and Eisner)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

- **2004 — annealing techniques** **(Smith and Eisner)**
- **2005 — contrastive estimation** **(Smith and Eisner)**
- **2006 — structural biasing** **(Smith and Eisner)**
- **2007 — common cover link representation** **(Seginer)**

State-of-the-Art: A Brief History

- **1992 — word classes** **(Carroll and Charniak)**
- **1998 — greedy linkage via mutual information** **(Yuret)**
- **2001 — iterative re-estimation with EM** **(Paskin)**
- **2004 — right-branching baseline**
  — valence (DMV) **(Klein and Manning)**

- **2004 — annealing techniques** **(Smith and Eisner)**
- **2005 — contrastive estimation** **(Smith and Eisner)**
- **2006 — structural biasing** **(Smith and Eisner)**
- **2007 — common cover link representation** **(Seginer)**
- **2008 — logistic normal priors** **(Cohen et al.)**

State-of-the-Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)
- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

- 2004 — annealing techniques (Smith and Eisner)
- 2005 — contrastive estimation (Smith and Eisner)
- 2006 — structural biasing (Smith and Eisner)
- 2007 — common cover link representation (Seginer)
- 2008 — logistic normal priors (Cohen et al.)
- 2009 — lexicalization and smoothing (Headden et al.)

State-of-the-Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)
- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

- 2004 — annealing techniques (Smith and Eisner)
- 2005 — contrastive estimation (Smith and Eisner)
- 2006 — structural biasing (Smith and Eisner)
- 2007 — common cover link representation (Seginer)
- 2008 — logistic normal priors (Cohen et al.)
- 2009 — lexicalization and smoothing (Headden et al.)
- 2009 — soft parameter tying (Cohen and Smith)

State-of-the-Art: Dependency Model with Valence

State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

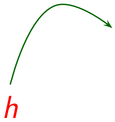
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

h

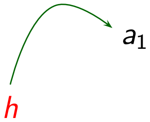
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



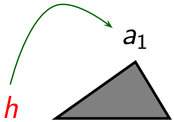
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



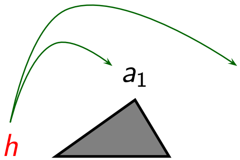
State-of-the-Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** (Klein and Manning, 2004)



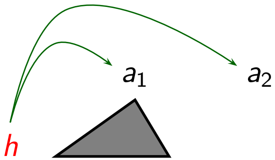
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



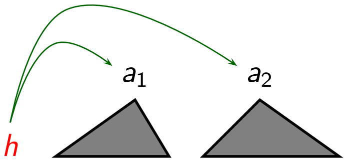
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



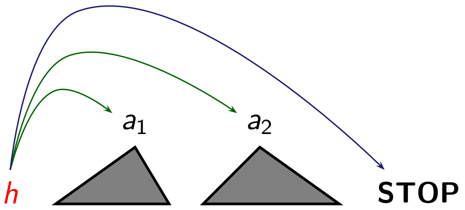
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



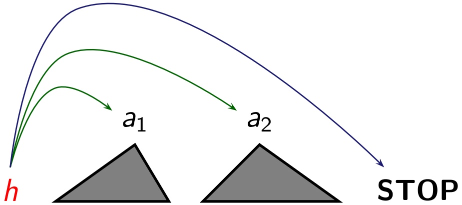
State-of-the-Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



State-of-the-Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence**/**adjacency** **(Klein and Manning, 2004)**



$$\mathbb{P}(t_h) = \prod_{dir \in \{\text{L},\text{R}\}} \left[\underline{\mathbb{P}_{\text{STOP}}(c_h, dir, \overbrace{1_{n=0}}^{adj})} \prod_{i=1}^{n} \underline{\mathbb{P}(t_{a_i})}\, \mathbb{P}_{\text{ATTACH}}(c_h, dir, c_{a_i}) \underline{(1 - \mathbb{P}_{\text{STOP}}(c_h, dir, \overbrace{1_{i=1}}^{adj}))} \right]_{n=|args(h,dir)|}$$

State-of-the-Art: Unsupervised Learning Engine

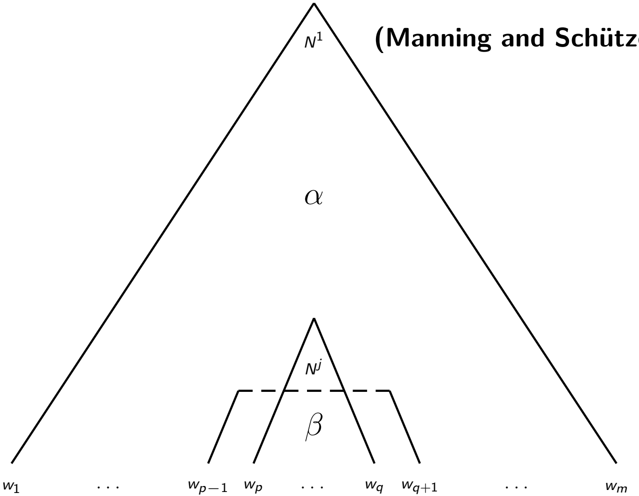
State-of-the-Art: Unsupervised Learning Engine

- **EM, via inside-outside re-estimation** (Baker, 1979)

State-of-the-Art: Unsupervised Learning Engine

- **EM, via inside-outside re-estimation (Baker, 1979)**

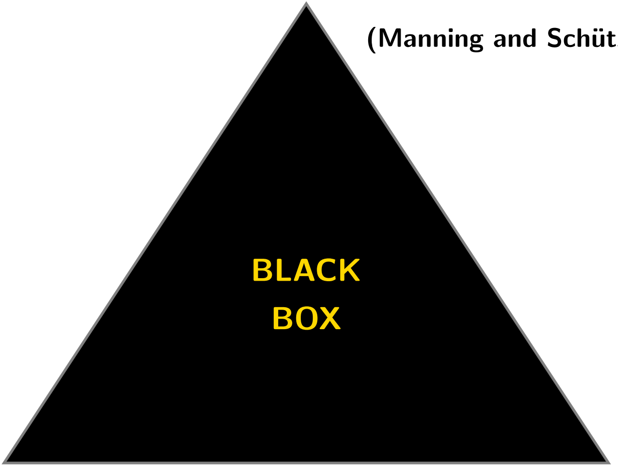
(Manning and Schütze, 1999)



State-of-the-Art: Unsupervised Learning Engine

- **EM, via inside-outside re-estimation (Baker, 1979)**

(Manning and Schütze, 1999)



w_1 $\cdots$ w_{p-1} w_p $\cdots$ w_q w_{q+1} $\cdots$ w_m

State-of-the-Art: The Standard Corpus

State-of-the-Art: The Standard Corpus

- **Training: WSJ10 (Klein, 2005)**

State-of-the-Art: The Standard Corpus

- **Training: WSJ10 (Klein, 2005)**
  - *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**

State-of-the-Art: The Standard Corpus

- **Training: WSJ10** **(Klein, 2005)**
  - *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**

State-of-the-Art: The Standard Corpus

- **Training: WSJ10 (Klein, 2005)**
  - *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... filtered down to sentences left with no more than 10 POS tags;**

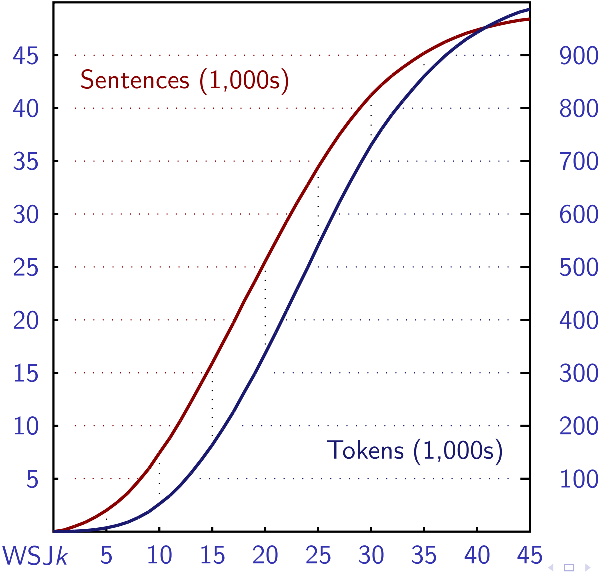
State-of-the-Art: The Standard Corpus

- **Training: WSJ10 (Klein, 2005)**
  - *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... filtered down to sentences left with no more than 10 POS tags;**
  - **... and converted to reference dependencies using "head percolation rules" (Collins, 1999).**

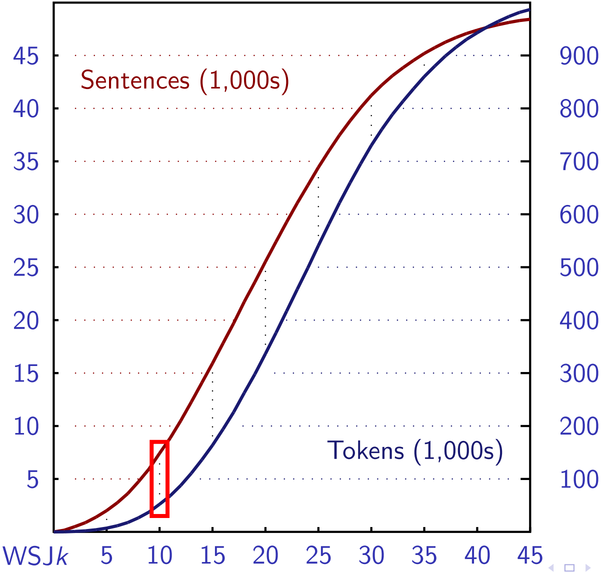
State-of-the-Art: The Standard Corpus

- **Training: WSJ10 (Klein, 2005)**
  - *The Wall Street Journal* **section of the Penn Treebank Project (Marcus et al., 1993)**
  - **... stripped of punctuation, etc.**
  - **... filtered down to sentences left with no more than 10 POS tags;**
  - **... and converted to reference dependencies using "head percolation rules" (Collins, 1999).**
- **Evaluation: Section 23 of WSJ$^{\infty}$ (all sentences).**

State-of-the-Art: The Standard Corpus



State-of-the-Art: The Standard Corpus



Issue I: Why so little data?

Issue I: Why so little data?

- **extra unlabeled data**
  helps semi-supervised parsing **(Suzuki et al., 2009)**

Issue I: Why so little data?

- **extra unlabeled data**
  helps semi-supervised parsing (Suzuki et al., 2009)

- **yet state-of-the-art unsupervised methods use even**
  less than what's available for supervised training...

Issue I: Why so little data?

- **extra unlabeled data**
  helps semi-supervised parsing (Suzuki et al., 2009)

- **yet state-of-the-art unsupervised methods use even less than what's available for supervised training...**

- **we will explore (three) judicious uses of data and simple, scalable machine learning techniques**

Issue II: Non-convex objective...

Issue II: Non-convex objective...

- **maximizing the probability of data (sentences):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

Issue II: Non-convex objective...

- **maximizing the probability of data (sentences):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

- **supervised objective would be convex (counting):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \sum_{s} \log \mathbb{P}_{\theta}(t^*(s)).$$

Issue II: Non-convex objective...

- **maximizing the probability of data (sentences):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

- **supervised objective would be convex (counting):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \sum_{s} \log \mathbb{P}_{\theta}(t^{*}(s)).$$

- **in general, $\hat{\theta}_{\text{SUP}} \neq \hat{\theta}_{\text{UNS}}$ and $\hat{\theta}_{\text{UNS}} \neq \tilde{\theta}_{\text{UNS}}$... (see CoNLL)**

Issue II: Non-convex objective...

- **maximizing the probability of data (sentences):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

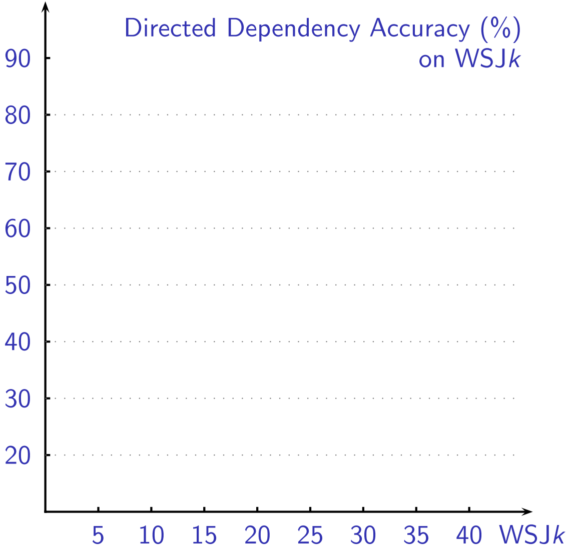
- **supervised objective would be convex (counting):**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \sum_{s} \log \mathbb{P}_{\theta}(t^*(s)).$$

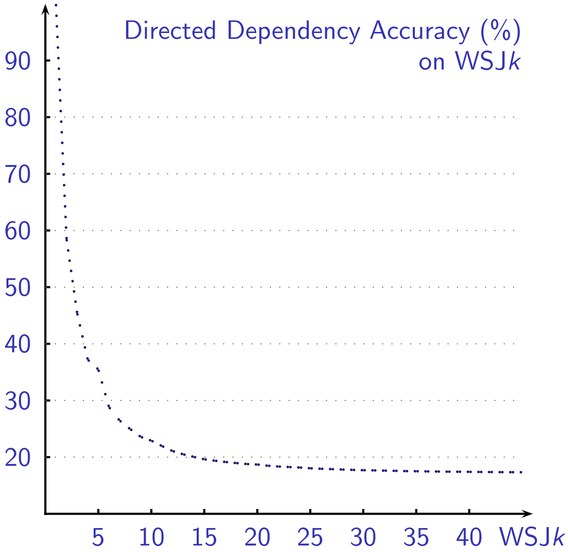
- **in general, $\hat{\theta}_{\text{SUP}} \neq \hat{\theta}_{\text{UNS}}$ and $\hat{\theta}_{\text{UNS}} \neq \tilde{\theta}_{\text{UNS}}$... (see CoNLL)**

- **initialization matters!**

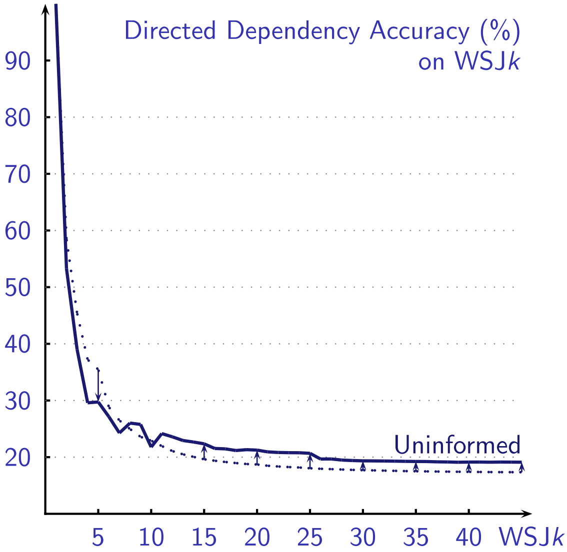
Issues: The Lay of the Land



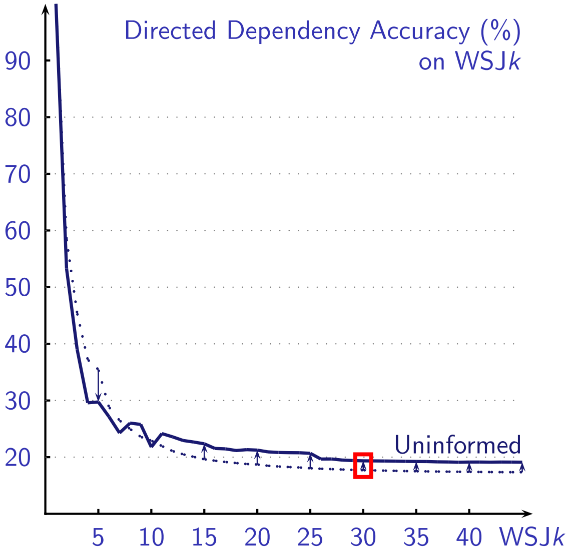
Issues: The Lay of the Land



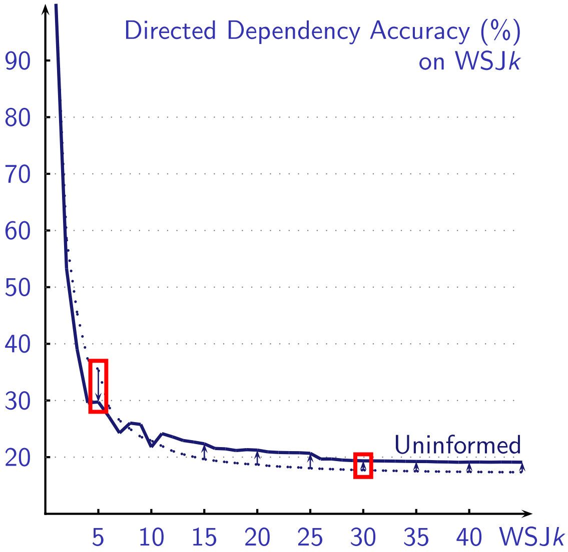
Issues: The Lay of the Land



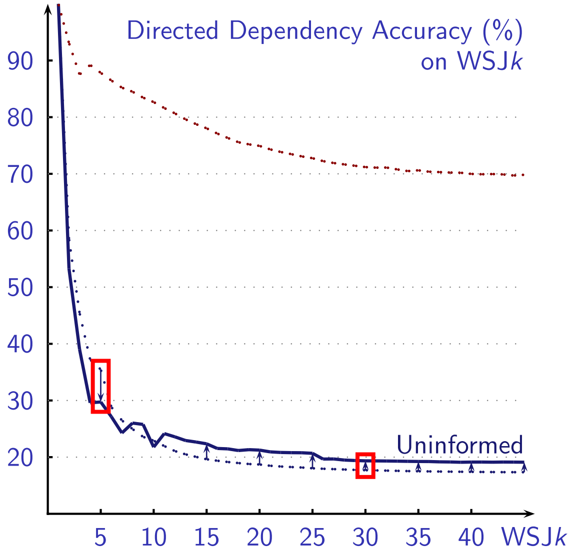
Issues: The Lay of the Land



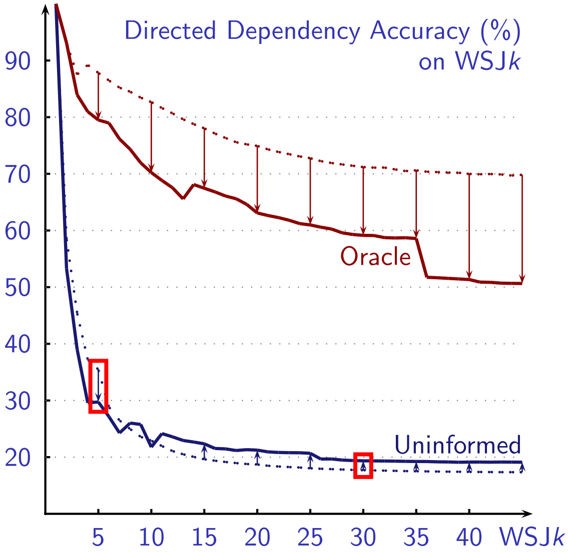
Issues: The Lay of the Land



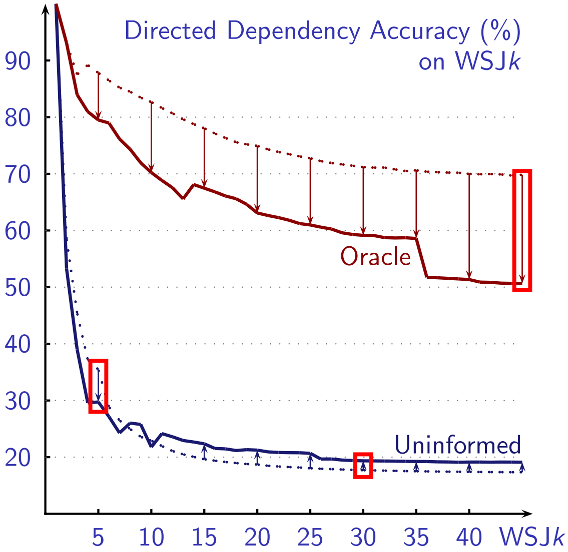
Issues: The Lay of the Land



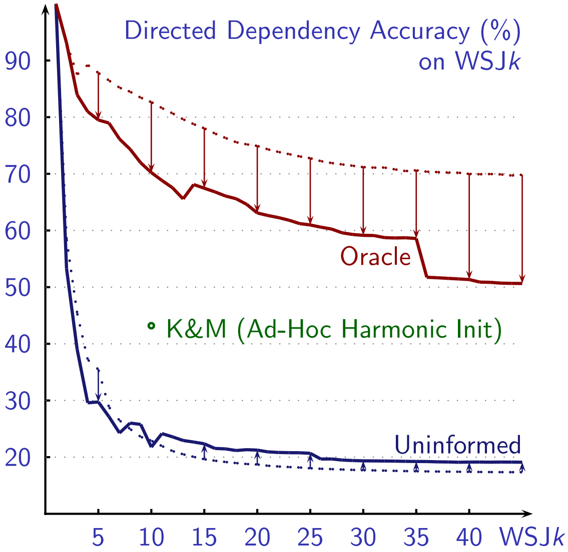
Issues: The Lay of the Land



Issues: The Lay of the Land



Issues: The Lay of the Land



Idea I: Baby Steps ... as Non-convex Optimization

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**

Idea I: Baby Steps ... as Non-convex Optimization

- global non-convex optimization is hard ...
- meta-heuristic: take guesswork out of local search
- start with an easy (convex) case
- slowly extend it to the fully complex target task

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

- **base case: sentences of length one (trivial — no init)**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

- **base case: sentences of length one (trivial — no init)**
- **incremental step: smooth WSJk; re-init WSJ$(k+1)$**

Idea I: Baby Steps ... as Non-convex Optimization

- **global non-convex optimization is hard ...**
- **meta-heuristic: take guesswork out of local search**
- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

- **base case: sentences of length one (trivial — no init)**
- **incremental step: smooth WSJk; re-init WSJ$(k+1)$**

- **... this *really is* grammar induction!**

Idea I: Baby Steps ... as Graduated Learning

Idea I: Baby Steps ... as Graduated Learning

- **WSJ1** — **Atone** **(verbs!)**

Idea I: Baby Steps ... as Graduated Learning

- $\boxed{\text{WSJ1}}$ — **Atone** **(verbs!)**

- $\boxed{\text{WSJ2}}$ — **Darkness fell.** **(nouns!)**
  It is.
  Judge Not

Idea I: Baby Steps ... as Graduated Learning

- $\boxed{\textbf{WSJ1}}$ — **Atone** (**verbs!**)
- $\boxed{\textbf{WSJ2}}$ — **Darkness fell.** (**nouns!**)
  It is.
  Judge Not
- $\boxed{\textbf{WSJ3}}$ — **Become a Lobbyist** (**determiners!**)
  But many have.
  They didn't.

Idea I: Baby Steps ... and Related Notions

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**
- **starting small** **(Elman, 1993)**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**
- **starting small** **(Elman, 1993)**
  - **scaffold on model complexity** **[restrict memory]**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**
- **starting small** **(Elman, 1993)**
  - **scaffold on model complexity** **[restrict memory]**
  - **scaffold on data complexity** **[restrict input]**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**
- **starting small** **(Elman, 1993)**
  - **scaffold on model complexity** **[restrict memory]**
  - **scaffold on data complexity** **[restrict input]**

controversy! **(Rohde and Plaut, 1999)**

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**
- **less is more** **(Kail, 1984; Newport, 1988; 1990)**
- **starting small** **(Elman, 1993)**
  - **scaffold on model complexity** **[restrict memory]**
  - **scaffold on data complexity** **[restrict input]**

controversy! **(Rohde and Plaut, 1999)**

- **stepping stones** **(Brown et al., 1993)**

Idea I: Baby Steps ... and Related Notions

- **shaping** (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]

controversy! (Rohde and Plaut, 1999)

- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** (Charniak and Johnson, 2005)

Idea I: Baby Steps ... and Related Notions

- **shaping** (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - scaffold on **model** complexity [restrict **memory**]
  - scaffold on **data** complexity [restrict **input**]

controversy! (Rohde and Plaut, 1999)

- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** (Charniak and Johnson, 2005)
- **curriculum learning** (Bengio et al., 2009)

Idea I: Baby Steps ... and Related Notions

- shaping (Skinner, 1938)
- less is more (Kail, 1984; Newport, 1988; 1990)
- starting small (Elman, 1993)
  - scaffold on model complexity [restrict memory]
  - scaffold on data complexity [restrict input]

controversy! (Rohde and Plaut, 1999)

- stepping stones (Brown et al., 1993)
- coarse-to-fine (Charniak and Johnson, 2005)
- curriculum learning (Bengio et al., 2009)
- continuation methods (Allgower and Georg, 1990)

Idea I: Baby Steps ... and Related Notions

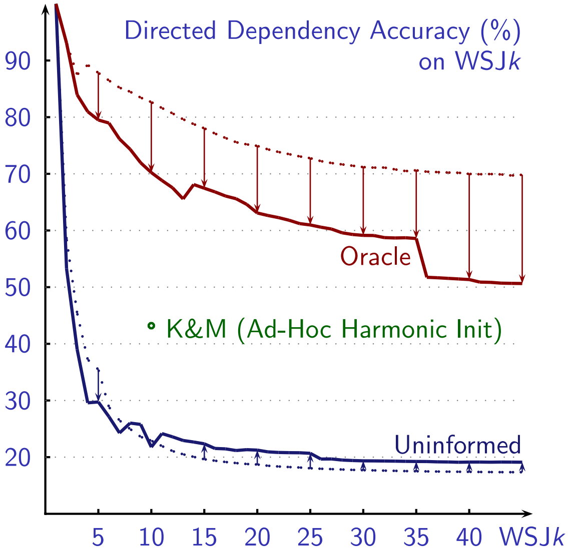
- **shaping** (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]

controversy! (Rohde and Plaut, 1999)

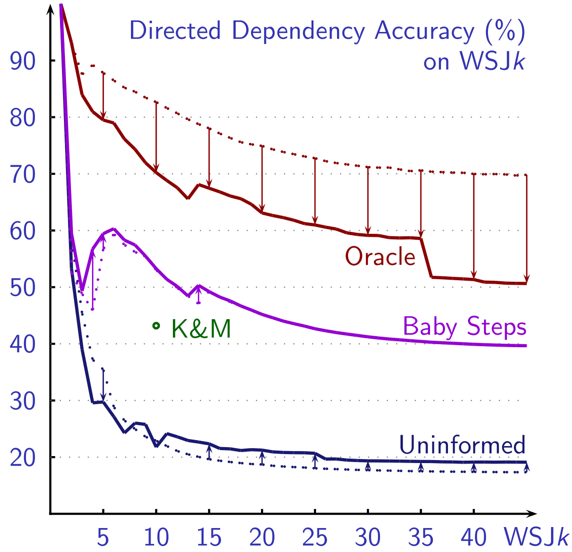
- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** (Charniak and Johnson, 2005)
- **curriculum learning** (Bengio et al., 2009)
- **continuation methods** (Allgower and Georg, 1990)

successive approximations!

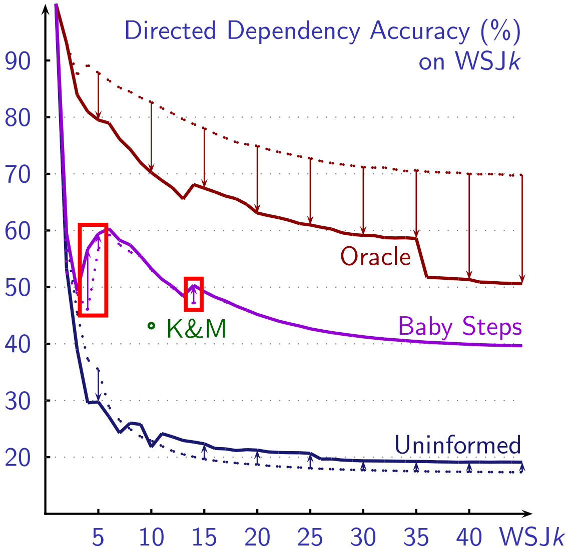
Idea I: Baby Steps ... Results!



Idea I: Baby Steps ... Results!



Idea I: Baby Steps ... Results!



Idea I: Baby Steps ... Concerns?

Idea I: Baby Steps ... Concerns?

- **ignores a good initializer**

Idea I: Baby Steps ... Concerns?

- **ignores a good initializer**
- **unnecessarily meticulous**

Idea I: Baby Steps ... Concerns?

- **ignores a good initializer**
- **unnecessarily meticulous**
- **excruciatingly slow!**

Idea I: Baby Steps ... Concerns?

- **ignores a good initializer**
- **unnecessarily meticulous**
- **excruciatingly slow!**

- **about a year behind state-of-the-art (on long sentences)**

Idea II: Less is More

Idea II: Less is More

- **short sentences are not representative (and few)**

Idea II: Less is More

- **short sentences are not representative (and few)**
- **long sentences are overwhelmingly difficult ...**

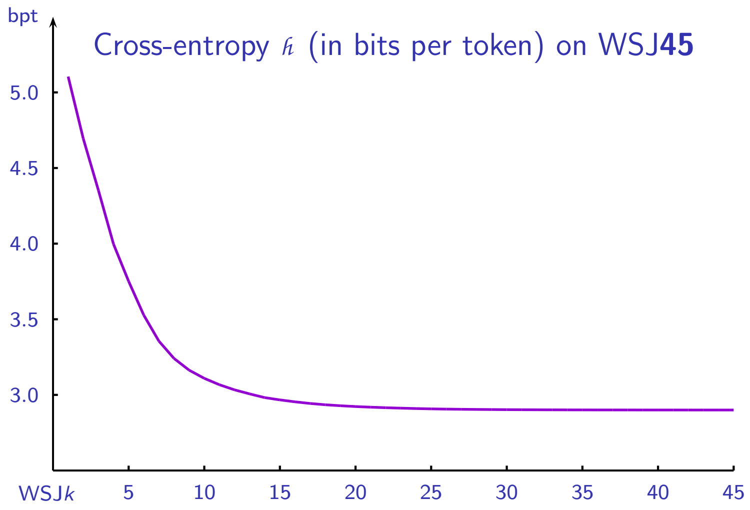
Idea II: Less is More

- **short sentences are not representative (and few)**
- **long sentences are overwhelmingly difficult ...**
- **is there a sweet spot data gradation?**

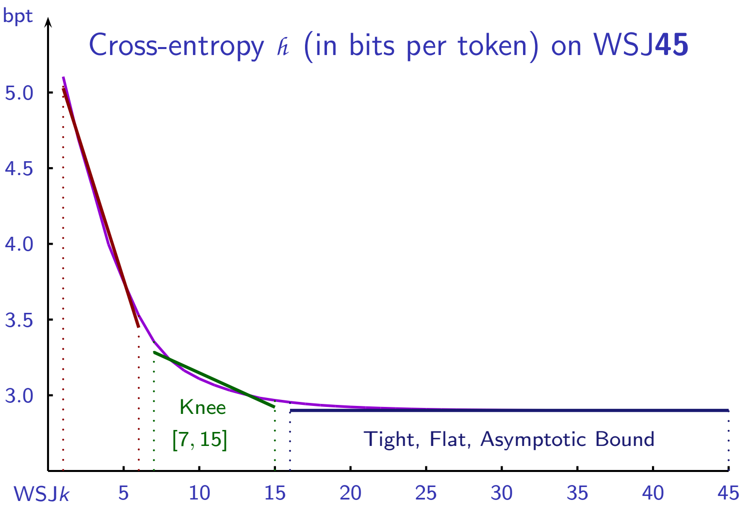
Idea II: Less is More

- **short sentences are not representative (and few)**
- **long sentences are overwhelmingly difficult ...**
- **is there a sweet spot data gradation?**
- **perhaps train where *Baby Steps* flatlines!**

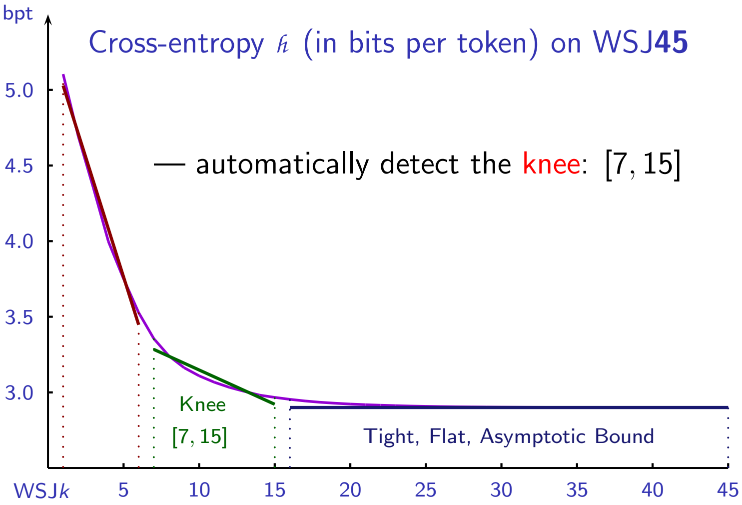
Idea II: Less is More ... the Learning Curve



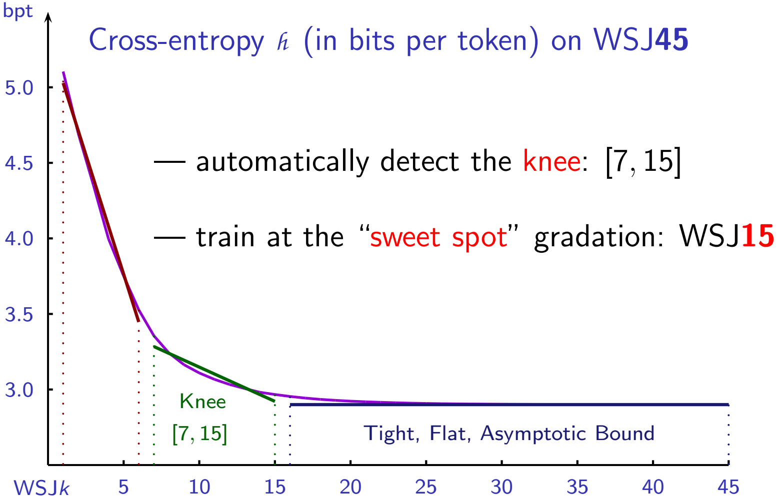
Idea II: Less is More ... the Learning Curve



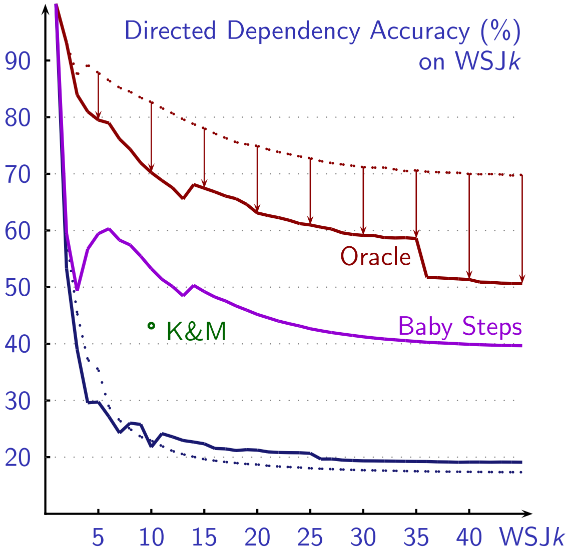
Idea II: Less is More ... the Learning Curve



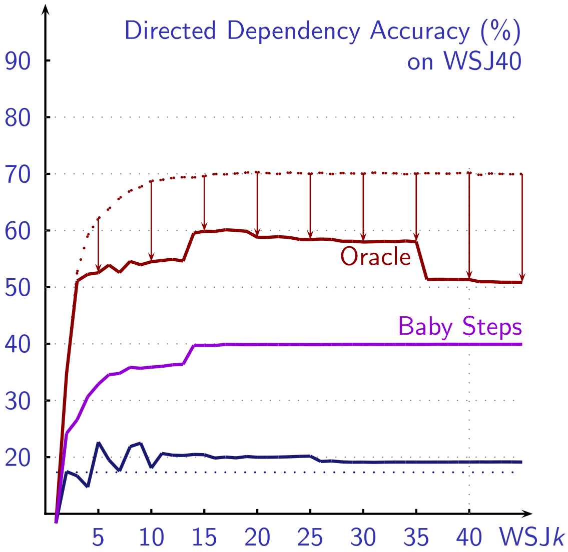
Idea II: Less is More … the Learning Curve



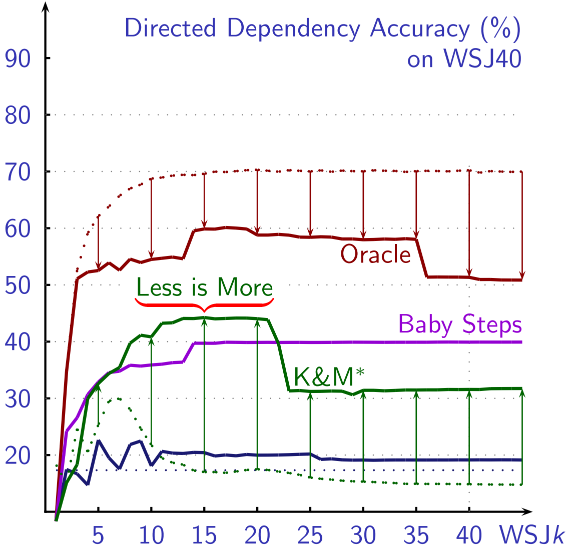
Idea II: Less is More ... Results!



Idea II: Less is More ... Results!



Idea II: Less is More ... Results!



Idea II: Less is More ... Concerns?

Idea II: Less is More ... Concerns?

- **discards most of the data**

Idea II: Less is More ... Concerns?

- **discards most of the data**

- **beats state-of-the-art (on long sentences, off WSJ15)**

Idea II: Less is More ... Concerns?

- **discards most of the data**
- **beats state-of-the-art (on long sentences, off WSJ15)**
- **ignores a decent complementary initialization strategy**

Idea III: Leapfrog ... a Hack

Idea III: Leapfrog ... a Hack

- **use *both* good systems!**

Idea III: Leapfrog ... a Hack

- **use *both* good systems!**

- **thorough training up to WSJ15, where it's cheap**

Idea III: Leapfrog ... a Hack

- **use *both* good systems!**

- **thorough training up to WSJ15, where it's cheap**
- **use both good initializers (mix their best parse trees)**

Idea III: Leapfrog ... a Hack

- **use** *both* **good systems!**

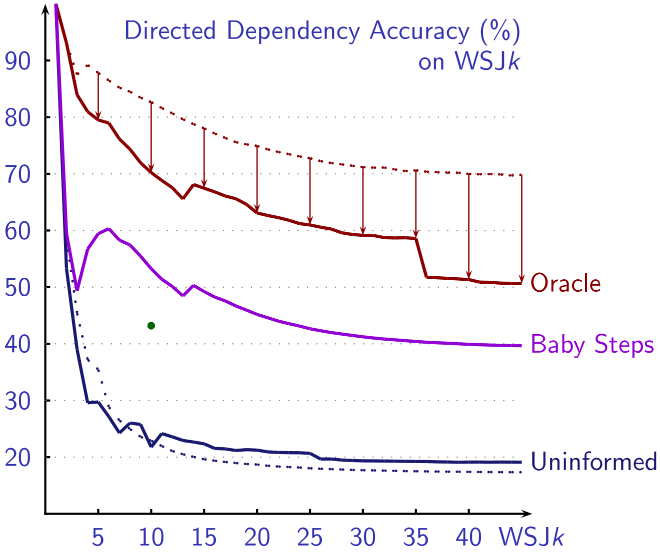
- **thorough training up to WSJ15, where it's cheap**
- **use both good initializers (mix their best parse trees)**
- **execute just a few steps of EM where it's expensive**

Idea III: Leapfrog ... a Hack

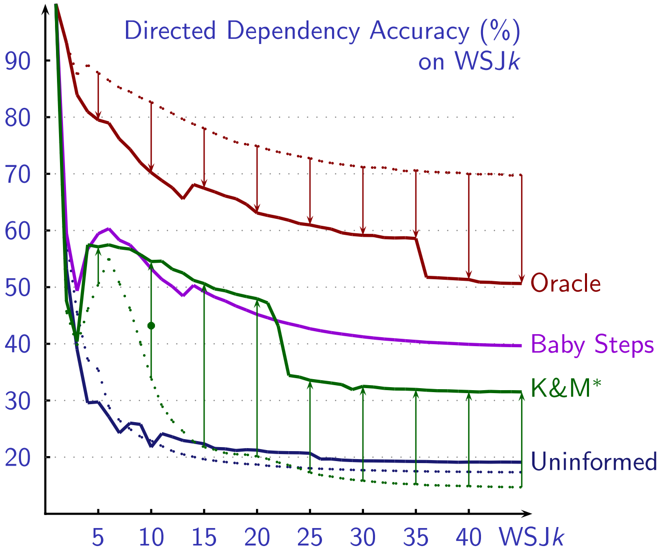
- **use *both* good systems!**

- **thorough training up to WSJ15, where it's cheap**
- **use both good initializers (mix their best parse trees)**
- **execute just a few steps of EM where it's expensive**
- **hop on from WSJ15 to WSJ45, via WSJ30...**

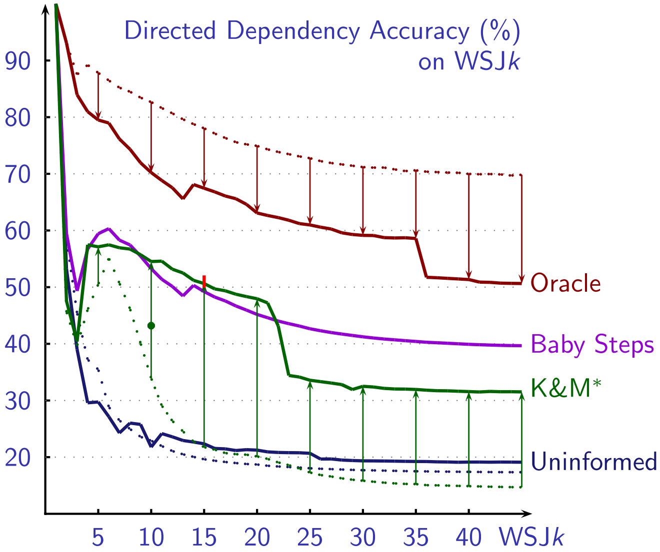
Idea III: Leapfrog ... Results!



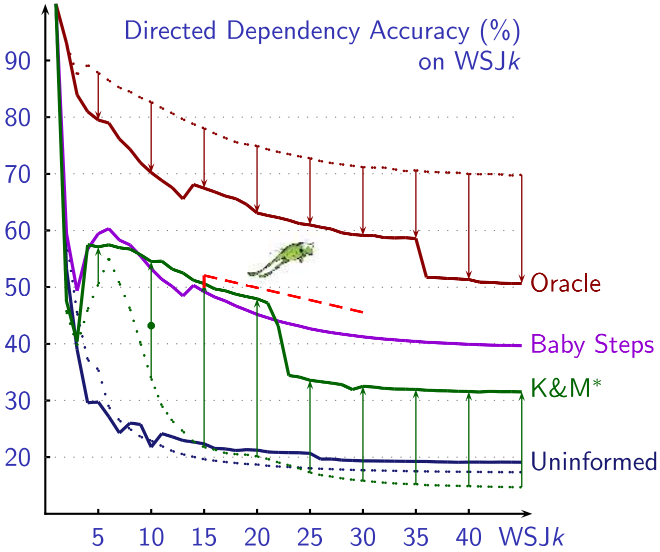
Idea III: Leapfrog ... Results!



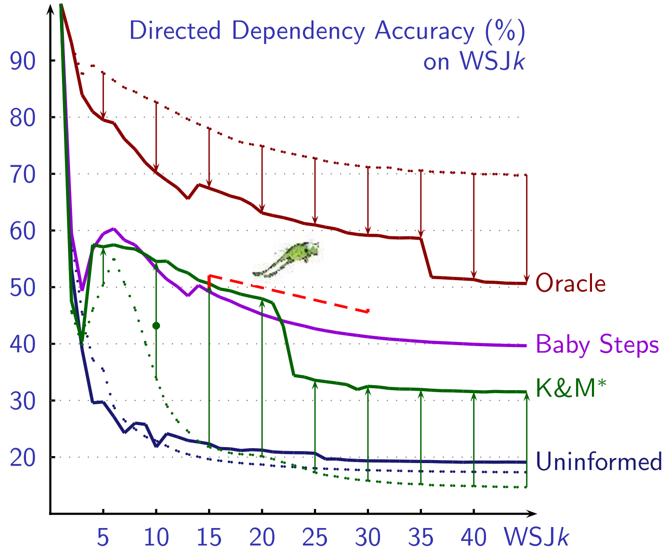
Idea III: Leapfrog ... Results!



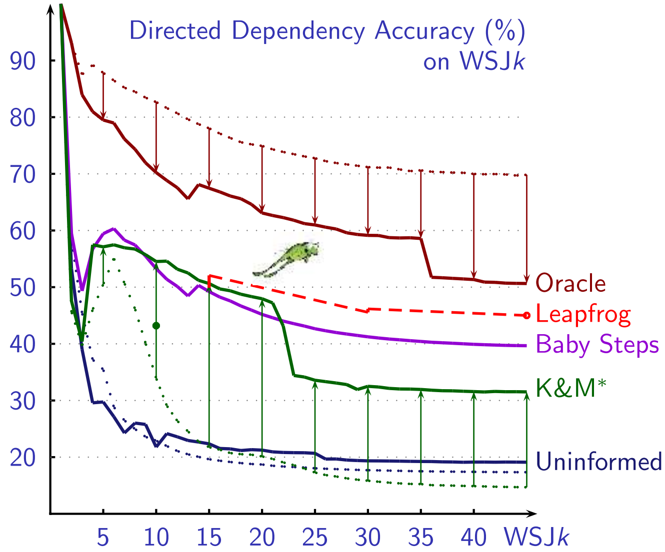
Idea III: Leapfrog ... Results!



Idea III: Leapfrog ... Results!



Idea III: Leapfrog ... Results!



Results: ... on Section 23 of WSJ

Results: ... on Section 23 of WSJ

Right-Branching	**(Klein and Manning, 2004)**	**31.7%**

Results: ... on Section 23 of WSJ

Right-Branching	**(Klein and Manning, 2004)**	**31.7%**
DMV	**@10**	**34.2%**

Results: ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)		**31.7%**
DMV		@10	**34.2%**
Baby Steps		@15	**39.2%**
Baby Steps		@45	**39.4%**

Results: ... on Section 23 of WSJ

Right-Branching		(Klein and Manning, 2004)	31.7%
DMV		@10	34.2%
Baby Steps		@15	39.2%
Baby Steps		@45	39.4%
Soft Parameter Tying		(Cohen and Smith, 2009)	42.2%

Results: ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%
Less is More	@15	44.1%

Results: ... on Section 23 of WSJ

Right-Branching		(Klein and Manning, 2004)	31.7%
DMV		@10	34.2%
Baby Steps		@15	39.2%
Baby Steps		@45	39.4%
Soft Parameter Tying		(Cohen and Smith, 2009)	42.2%
Less is More		@15	44.1%
Leapfrog		@45	45.0%

Summary

Summary

- **explored scaffolding on data complexity**

Summary

- **explored scaffolding on data complexity**
- **awareness of data complexity does help!**

Summary

- **explored scaffolding on data complexity**
- **awareness of data complexity does help!**
- **beats state-of-the-art with older techniques**

Conclusion

Conclusion

- **(need a less adversarial learning algorithm)**

Conclusion

- (need a less adversarial learning algorithm)
- paradox: improved performance with less data

Conclusion

- (need a less adversarial learning algorithm)

- paradox: improved performance with less data
- despite discarding samples from the true (test) distribution

Conclusion

- **(need a less adversarial learning algorithm)**

- **paradox: improved performance with less data**
- **despite discarding samples from the true (test) distribution**

- **focusing on simple examples guides unsupervised learning**

Conclusion

- (need a less adversarial learning algorithm)

- paradox: improved performance with less data
- despite discarding samples from the true (test) distribution

- focusing on simple examples guides unsupervised learning
- mirrors supervised boosting (Freund and Schapire, 1997)

Teaser

Teaser

- **we push the state-of-the-art further, to 50.4% (up another 5%) using even faster and simpler methods!**

Teaser

- **we push the state-of-the-art further, to 50.4% (up another 5%) using even faster and simpler methods!**

- **... hear us at CoNLL and ACL (Spitkovsky et al., 2010)**

Teaser

- we push the state-of-the-art further, to 50.4% (up another 5%) using even faster and simpler methods!

- ... hear us at CoNLL and ACL (Spitkovsky et al., 2010)

- similar approaches may apply in other settings (e.g., word alignment)

Teaser

- we push the state-of-the-art further, to 50.4% (up another 5%) using even faster and simpler methods!

- ... hear us at CoNLL and ACL (Spitkovsky et al., 2010)

- similar approaches may apply in other settings (e.g., word alignment)

- ... more to come!

Thanks!

Questions?