Punctuation: Making a Point in Unsupervised Dependency Parsing

Valentin I. Spitkovsky

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

Example: Raw Word Stream

Example: Raw Word Stream

ALTHOUGH IT PROBABLY HAS REDUCED THE LEVEL OF EXPENDITURES FOR SOME PURCHASERS UTILIZATION MANAGEMENT LIKE MOST OTHER COST CONTAINMENT STRATEGIES DOESN'T APPEAR TO HAVE ALTERED THE LONG-TERM RATE OF INCREASE IN HEALTH-CARE COSTS THE INSTITUTE OF MEDICINE AN AFFILIATE OF THE NATIONAL ACADEMY OF SCIENCES CONCLUDED AFTER A TWO-YEAR STUDY

Example:

Example:

- **formatting (missing structural cues):**

Example:

- **formatting (missing structural cues):**
  — e.g., punctuation and capitalization

Example:

- **formatting (missing structural cues):**
  — e.g., punctuation and capitalization

- **raw word streams often difficult even for humans**
  — e.g., transcribed utterances (Kim and Woodland, 2002)

Example:

IN PRP RB VBZ VBN DT NN IN NNS IN DT NNS NN NN IN RBS JJ NN NN NNS VBZ RB VB TO VB VBN DT JJ NN IN NN IN JJ NNS DT NNP IN NNP DT NN IN DT NNP NNP IN NNPS VBD IN DT JJ NN

Example:

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**],

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] —

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] — [$_{\text{PP}}$ **like most other cost containment strategies**] —

Example:

[$_{SBAR}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{NP}$ **utilization management**] — [$_{PP}$ **like most other cost containment strategies**] — [$_{VP}$ **doesn't appear to have altered the long-term rate of increase in health-care costs**],

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] — [$_{\text{PP}}$ **like most other cost containment strategies**] — [$_{\text{VP}}$ **doesn't appear to have altered the long-term rate of increase in health-care costs**], [$_{\text{NP}}$ **the Institute of Medicine**],

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] — [$_{\text{PP}}$ **like most other cost containment strategies**] — [$_{\text{VP}}$ **doesn't appear to have altered the long-term rate of increase in health-care costs**], [$_{\text{NP}}$ **the Institute of Medicine**], [$_{\text{NP}}$ **an affiliate of the National Academy of Sciences**],

Example:

[$_{\text{SBAR}}$ Although it probably has reduced the level of expenditures for some purchasers], [$_{\text{NP}}$ utilization management] — [$_{\text{PP}}$ like most other cost containment strategies] — [$_{\text{VP}}$ doesn't appear to have altered the long-term rate of increase in health-care costs], [$_{\text{NP}}$ the Institute of Medicine], [$_{\text{NP}}$ an affiliate of the National Academy of Sciences], [$_{\text{VP}}$ concluded after a two-year study].

Intuition:

Intuition:

- **punctuation is a strong structural cue**

Intuition:

- **punctuation is a strong structural cue — demarcates separable fragments**

Intuition:

- **punctuation is a strong structural cue — demarcates separable fragments**

- **we will make simplifying independence assumptions**

Intuition:

- **punctuation is a strong structural cue**
  — demarcates separable fragments

- **we will make simplifying independence assumptions**
  — (unreasonably) strong in training

Intuition:

- punctuation is a strong structural cue
  — demarcates separable fragments

- we will make simplifying independence assumptions
  — (unreasonably) strong in training

- less crude in inference
  — (reasonably) weak in final decoding

Intuition:

- **strong constraint**

Intuition:

- **strong constraint: (head $\leftarrow$ head) in training**

Intuition:

- **strong constraint: (head $\leftarrow$ head) in training**

word head , head word word ,

head word word word word word word word .

Intuition:

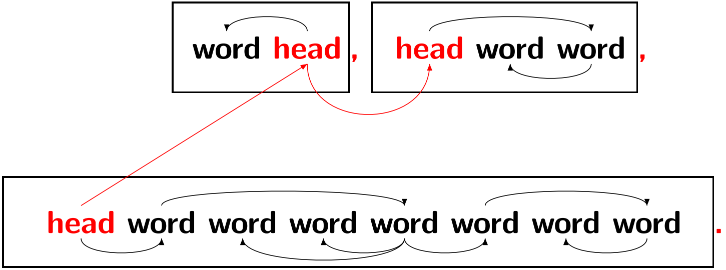
- **strong constraint: (head $\leftarrow$ head) in training**

word head, **head word word**,

head word word word word word word word .

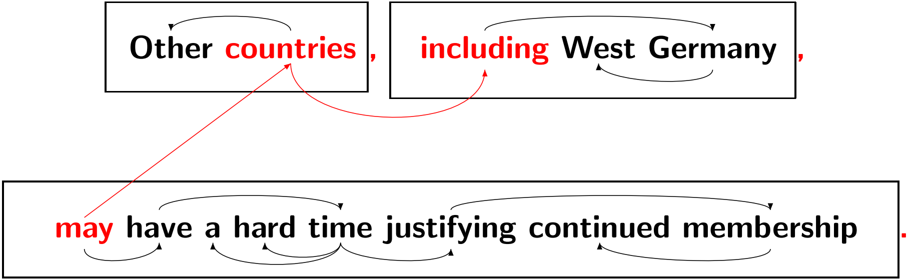
Intuition:

- **strong constraint: (head ← head) in training**



Intuition:

- strong constraint: (head $\leftarrow$ head) in training



Intuition:

- weak constraint

Intuition:

- **weak constraint: (head $\leftarrow$ external word) in inference**

Intuition:

- **weak constraint: (head $\leftarrow$ external word) in inference**

word word head word word word

,

head word word word word word word word

.

Intuition:

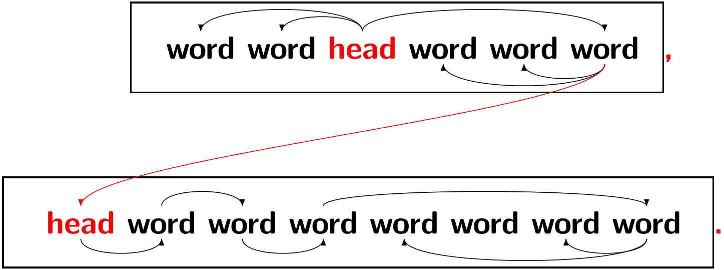
- **weak constraint: (head $\leftarrow$ external word) in inference**

word word **head** word word word ,

head word word word word word word word .

Intuition:

- **weak constraint: (head ← external word) in inference**



Intuition:

- **weak constraint: (head $\leftarrow$ external word) in inference**

IFI also has nonvoting preferred shares ,

which are quoted on the Milan stock exchange .

Linguistic Analysis:

- **punctuation and syntax are related**
  (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998, ***inter alia*****)**

Linguistic Analysis:

- **punctuation and syntax are related**
  (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998, *inter alia*)

- **49.4% of inter-punctuation fragments are constituents**

Linguistic Analysis:

- **punctuation and syntax are related**
  (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998, *inter alia*)

- **49.4% of inter-punctuation fragments are constituents**

- **lowest dominating non-terminals:**

	%
S	32.5
NP	27.2
VP	13.3
PP	10.1
SBAR	6.7
ADVP	3.3
QP	2.5
SINV	2.0
ADJP	1.0
	98.5

Linguistic Analysis:

Linguistic Analysis:

- **strong (in training)**

Linguistic Analysis:

- **strong** (in training), e.g.,

... arrests followed a “ *Snake Day* ” at Utrecht ...

Linguistic Analysis:

- **strong (in training), e.g.,**

… arrests followed a " *Snake Day* " at Utrecht …

— already 74.0% agreement with head-percolated trees

Linguistic Analysis:

Linguistic Analysis:

- **weak (in inference)**

Linguistic Analysis:

- **weak** (in inference), e.g.,

Maryland Club also distributes tea , which ...

Linguistic Analysis:

- **weak (in inference), e.g.,**

Maryland Club also distributes tea , which ...

— now 92.9% agreement with head-percolated trees

Linguistic Analysis:

Linguistic Analysis:

- **generalization:**

Linguistic Analysis:

- **generalization**:
  — no path from the root may enter a fragment twice

Linguistic Analysis:

- **generalization:**
  — no path from the root may enter a fragment twice
  — 95.0% agreement with head-percolated trees

Linguistic Analysis:

- **generalization:**
  — no path from the root may enter a fragment twice
  — 95.0% agreement with head-percolated trees

- **simple violations:**

Linguistic Analysis:

- **generalization:**
  — no path from the root may enter a fragment twice
  — 95.0% agreement with head-percolated trees

- **simple violations: "seamless" quotations**

Her recent report classifies the stock as a "hold."

Linguistic Analysis:

- **generalization:**
  — no path from the root may enter a fragment twice
  — 95.0% agreement with head-percolated trees

- **simple violations: "seamless" quotations and even lists**

Her recent report classifies the stock as a "hold."

The company said its directors, management and subsidiaries will remain long-term investors and ...

Motivation: "Profiting from Markup"

Motivation: "Profiting from Markup"

..., whereas McCain is secure on the topic, Obama <a>[$_{VP}$ worries about winning the pro-Israel vote].

Motivation: "Profiting from Markup"

..., whereas McCain is secure on the topic, Obama <a>[$_{\text{VP}}$ worries about winning the pro-Israel vote].

"Capitalizing on Punctuation"

Motivation: "Profiting from Markup"

..., whereas McCain is secure on the topic, Obama <a>[$_{VP}$ worries about winning the pro-Israel vote].

"Capitalizing on Punctuation"
— more common (particularly in long sentences)

Motivation: "Profiting from Markup"

..., whereas McCain is secure on the topic, Obama <a>[$_{VP}$ worries about winning the pro-Israel vote].

"Capitalizing on Punctuation"

— more common (particularly in long sentences)
— more uniform (better coverage of constructs)

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 ...

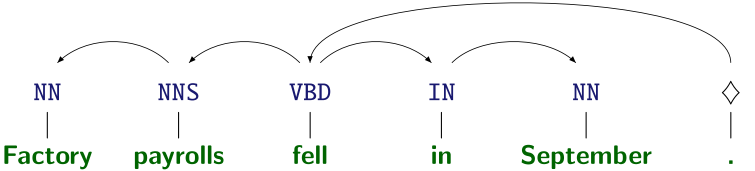
NN	NNS	VBD	IN	NN	♢
Factory	payrolls	fell	in	September	.

Problem: Unsupervised Learning of Parsing

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

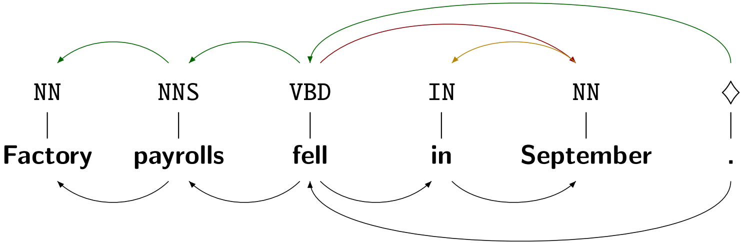
  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

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

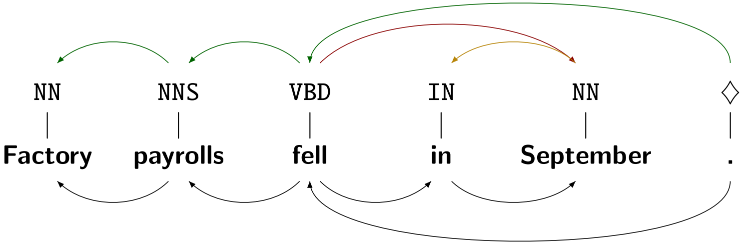


Scoring: Directed Dependency Accuracy

Scoring: Directed Dependency Accuracy

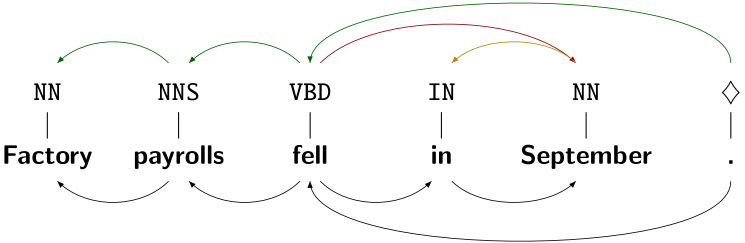


Scoring: Directed Dependency Accuracy



Directed score: $\frac{3}{5} = 60\%$

Scoring: Directed Dependency Accuracy



Directed score: $\frac{3}{5} = 60\%$ **(right/left-branching baselines:** $\frac{2}{5} = 40\%$**).**

DMV: Dependency Model with Valence

DMV: Dependency Model with Valence

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

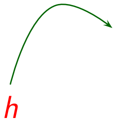
DMV: Dependency Model with Valence

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

h

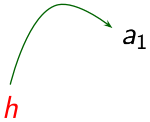
DMV: Dependency Model with Valence

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



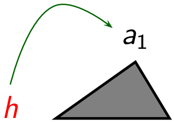
DMV: Dependency Model with Valence

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



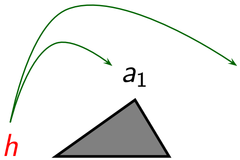
DMV: Dependency Model with Valence

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



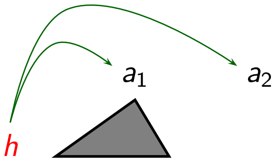
DMV: Dependency Model with Valence

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



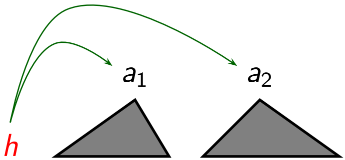
DMV: Dependency Model with Valence

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



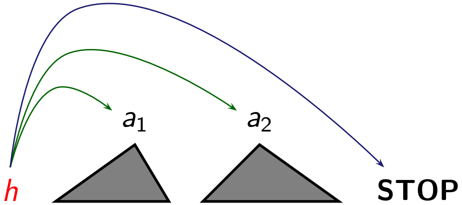
DMV: Dependency Model with Valence

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



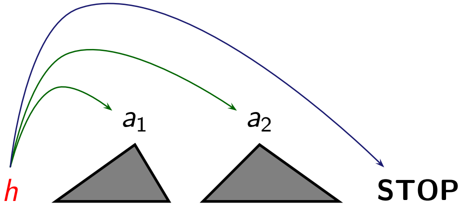
DMV: Dependency Model with Valence

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



DMV: Dependency Model with Valence

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



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

Learning: Viterbi EM

Learning: Viterbi EM

- **well-suited to long sentences, which are more punctuation-rich**

(Spitkovsky et al., CoNLL 2010)

Learning: Viterbi EM

- **well-suited to long sentences, which are more punctuation-rich**

  (Spitkovsky et al., CoNLL 2010)

- **fast, simple and easily admits constraints**

  (Spitkovsky et al., ACL 2010)

Constraints: Parser Induction

- **the model, i.e., projective trees (Klein and Manning, 2004) — Dependency Model with Valence (DMV)**

Constraints: Parser Induction

- **the model, i.e., projective trees** **(Klein and Manning, 2004)**
  — Dependency Model with Valence (DMV)

```
(((List (the fares (for ((flight) (number 891)))))) .)
```

- **partial bracketings** **(Pereira and Schabes, 1992)**

Constraints: Parser Induction

- **the model, i.e., projective trees** (Klein and Manning, 2004)
  — Dependency Model with Valence (DMV)

```
(((List (the fares (for ((flight) (number 891)))))) .)
```

- **partial bracketings** (Pereira and Schabes, 1992)
  - synchronous grammars (Alshawi and Douglas, 2000)
  - linear-time parsing (Seginer, 2007)
  - skewness of trees (Seginer, 2007)
  - Zipfian distribution of words (Seginer, 2007)
  - sparse posterior regularization (Ganchev et al., 2009)
  - web markup-induced constraints (Spitkovsky et al., 2010)
  - semantic cues (Naseem and Barzilay, 2011)

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^{\infty}$

(Section 23, all sentences)

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^\infty$ (Section 23, all sentences)
Standard Training	**52.0**

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^\infty$	**(Section 23, all sentences)**
Punctuation as Words	**41.7**	**(-10.3)**
Standard Training	**52.0**	

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^{\infty}$	**(Section 23, all sentences)**
Punctuation as Words	**41.7**	**(-10.3)**
Standard Training	**52.0**	
w/Constrained Inference	**54.0**	**(+2.0)**

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ^{∞}	(Section 23, all sentences)
Punctuation as Words	**41.7**	**(-10.3)**
Standard Training	**52.0**	
w/Constrained Inference	**54.0**	**(+2.0)**
Constrained Training	**55.6**	**(+3.6)**

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ^{∞}	(Section 23, all sentences)
Punctuation as Words	**41.7**	**(-10.3)**
Standard Training	**52.0**	
w/Constrained Inference	**54.0**	**(+2.0)**
Constrained Training	**55.6**	**(+3.6)**
w/Constrained Inference	**57.4**	**(+1.8)**

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^\infty$	**(Section 23, all sentences)**
Punctuation as Words	**41.7**	**(-10.3)**
Standard Training	**52.0**	
w/Constrained Inference	**54.0**	**(+2.0)**
Constrained Training	**55.6**	**(+3.6)**
w/Constrained Inference	**57.4**	**(+1.8)**
Supervised DMV	**69.8**	

Experimental Results: Unlexicalized

- **directed dependency accuracies for baselines, inference, training and an oracle:**

	WSJ$^{\infty}$	(Section 23, all sentences)
Punctuation as Words	41.7	(-10.3)
Standard Training	52.0	
w/Constrained Inference	54.0	(+2.0)
Constrained Training	55.6	(+3.6)
w/Constrained Inference	57.4	(+1.8)
Supervised DMV	69.8	
w/Constrained Inference	73.0	(+3.2)

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	WSJ$^\infty$

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	WSJ$^{\infty}$
Unlexicalized	**57.4**

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	$\textbf{WSJ}^{\infty}$
(Spitkovsky et al., ACL 2010)	**50.4**
Unlexicalized	**57.4**

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	WSJ$^{\infty}$
(Spitkovsky et al., ACL 2010)	**50.4**
Lexicalized **(Gillenwater et al., 2010)**	**53.3**
Unlexicalized	**57.4**

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	WSJ$^{\infty}$
(Spitkovsky et al., ACL 2010)	**50.4**
Lexicalized **(Gillenwater et al., 2010)**	**53.3**
Lexicalized **(Blunsom and Cohn, 2010)**	**55.7**
Unlexicalized	**57.4**

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	$\mathbf{WSJ}^{\infty}$
(Spitkovsky et al., ACL 2010)	**50.4**
Lexicalized **(Gillenwater et al., 2010)**	**53.3**
Lexicalized **(Blunsom and Cohn, 2010)**	**55.7**
Unlexicalized	**57.4**
Lexicalized Constrained Training	**58.0**

Experimental Results: Lexicalized

- **directed dependency accuracies compared to previous state-of-the-art**

	$\mathbf{WSJ}^{\infty}$
(Spitkovsky et al., ACL 2010)	**50.4**
Lexicalized **(Gillenwater et al., 2010)**	**53.3**
Lexicalized **(Blunsom and Cohn, 2010)**	**55.7**
Unlexicalized	**57.4**
Lexicalized Constrained Training	**58.0**
w/Constrained Infernce	**58.4**

Experimental Results: "Fully" Unsupervised

Experimental Results: "Fully" Unsupervised

- **constraints sufficiently strong to abandon gold tags**

Experimental Results: "Fully" Unsupervised

- **constraints sufficiently strong to abandon gold tags**

	$\mathbf{WSJ}^{\infty}$
(this work)	**58.4**

Experimental Results: "Fully" Unsupervised

- **constraints sufficiently strong to abandon gold tags**

	$\mathbf{WSJ^{\infty}}$
(this work)	**58.4**
w/o Gold Tags	**58.2**

Experimental Results: "Fully" Unsupervised

- **constraints sufficiently strong to abandon gold tags**

	WSJ$^\infty$
(this work)	**58.4**
w/o Gold Tags	**58.2**

- **using Clark's (2000) unsupervised clusters**

Experimental Results: "Fully" Unsupervised

- **constraints sufficiently strong to abandon gold tags**

	WSJ$^\infty$
(this work)	58.4
w/o Gold Tags	58.2

- **using Clark's (2000) unsupervised clusters — constructed by Finkel and Manning (2009) for NER**

`http://nlp.stanford.edu/software/`
`stanford-postagger-2008-09-28.tar.gz:`
`models/egw.bnc.200`

Experimental Results: “Fully” Unsupervised

- **constraints sufficiently strong to abandon gold tags**

	WSJ^{∞}
(this work)	**58.4**
w/o Gold Tags	**58.2**

- **using Clark’s (2000) unsupervised clusters — constructed by Finkel and Manning (2009) for NER**

`http://nlp.stanford.edu/software/`
`stanford-postagger-2008-09-28.tar.gz:`
`models/egw.bnc.200`

(Come see our poster at EMNLP!)

Experimental Results: Multi-Lingual

- **further evaluation against CoNLL 2006/7 data sets — results generalize across languages:**

Arabic	**2006**		
	'7		
Basque	**'7**		
Bulgarian	**'6**		
Catalan	**'7**		
Czech	**'6**		
	'7		
Danish	**'6**		
Dutch	**'6**		
English	**'7**		
German	**'6**		
Greek	**'7**		
Hungarian	**'7**		
Italian	**'7**		
Japanese	**'6**		
Portuguese	**'6**		
Slovenian	**'6**		
Spanish	**'6**		
Swedish	**'6**		
Turkish	**'6**		
	'7		
Average:			

Experimental Results: Multi-Lingual

- **further evaluation against CoNLL 2006/7 data sets — results generalize across languages:**

		Inference Only	
Arabic	2006	+0.1	
	'7	+0.9	
Basque	'7	+0.8	
Bulgarian	'6	+1.1	
Catalan	'7	+0.8	
Czech	'6	+0.9	
	'7	+1.0	
Danish	'6	+0.9	
Dutch	'6	+1.0	
English	'7	+1.3	
German	'6	+0.8	
Greek	'7	+0.5	
Hungarian	'7	+0.4	
Italian	'7	+0.1	
Japanese	'6	+0.0	
Portuguese	'6	+0.7	
Slovenian	'6	+2.0	
Spanish	'6	+0.8	
Swedish	'6	+0.5	
Turkish	'6	+0.1	
	'7	+0.2	
Average:		+0.7	

Experimental Results: Multi-Lingual

- further evaluation against CoNLL 2006/7 data sets — results generalize across languages:

		Inference Only	Training & Inference
Arabic	2006	+0.1	+1.1
	'7	+0.9	+2.6
Basque	'7	+0.8	+0.6
Bulgarian	'6	+1.1	+1.6
Catalan	'7	+0.8	+0.9
Czech	'6	+0.9	+3.0
	'7	+1.0	+2.7
Danish	'6	+0.9	+0.2
Dutch	'6	+1.0	+3.0
English	'7	+1.3	+2.8
German	'6	+0.8	+1.6
Greek	'7	+0.5	+0.7
Hungarian	'7	+0.4	+1.4
Italian	'7	+0.1	-0.8
Japanese	'6	+0.0	+0.1
Portuguese	'6	+0.7	+0.8
Slovenian	'6	+2.0	+2.8
Spanish	'6	+0.8	+0.8
Swedish	'6	+0.5	+0.8
Turkish	'6	+0.1	+1.0
	'7	+0.2	+0.1
Average:		**+0.7**	**+1.3**

Thoughts:

Thoughts:

- **extend existing parsers**

Thoughts:

- **extend existing parsers**
  — no need to retrain models

Thoughts:

- **extend existing parsers**
  — no need to retrain models
  — supervised systems?

Thoughts:

- extend existing parsers
  — no need to retrain models
  — supervised systems?

- would prosody aid with induction from speech?

Thoughts:

- extend existing parsers
  — no need to retrain models
  — supervised systems?

- would prosody aid with induction from speech?
  — "as words" breaks *n*-grams (Kahn et al., 2005)

Summary:

Summary:

- **punctuation helps dependency grammar induction**

Summary:

- **punctuation helps dependency grammar induction — even better than markup...**

Summary:

- **punctuation helps dependency grammar induction — even better than markup...**

- **a popular approach: powerful models**

Summary:

- **punctuation helps dependency grammar induction — even better than markup...**

- **a popular approach: powerful models — priors prevent overfitting**

Summary:

- punctuation helps dependency grammar induction
  — even better than markup...
- a popular approach: powerful models
  — priors prevent overfitting
- an alternative: overly simple models

Summary:

- **punctuation helps dependency grammar induction — even better than markup...**
- **a popular approach: powerful models — priors prevent overfitting**
- **an alternative: overly simple models — constraints prevent *under*fitting**

Thanks!

Punctuation. It works...

Any questions?