Profiting from Mark-Up:
Hyper-Text Annotations
for Guided Parsing

Valentin I. Spitkovsky

with Daniel Jurafsky (Stanford University)
and Hiyan Alshawi (Google Inc.)
Constraints: Supervised and Unsupervised
Constraints: Supervised and Unsupervised

- compact summaries of high-level insights into a domain
Constraints: Supervised and Unsupervised

- compact summaries of high-level insights into a domain — e.g., a sentence should have a verb
Constraints: Supervised and Unsupervised

- compact summaries of high-level insights into a domain
  — e.g., a sentence should have a verb
  — can significantly reduce the search space
Constraints: **Supervised** and **Unsupervised**

- **Compact summaries** of high-level insights into a domain — e.g., a sentence should have a verb
- can significantly **reduce** the **search space**
- easier to list than annotating data (a few key rules)
Constraints: **Supervised** and **Unsupervised**

- **compact summaries** of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly **reduce** the **search space**
  - easier to list than annotating data (a few key rules)
  - enforce rather than model (avoid non-local features)
Constraints: **Supervised** and **Unsupervised**

- **compact summaries** of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly **reduce** the search space
  - easier to list than annotating data (a few key rules)
  - enforce rather than model (avoid non-local features)
  - enable simple, painless NLP, e.g., **joint inference**
Constraints: **Supervised and Unsupervised**

- compact summaries of high-level insights into a domain — e.g., a sentence should have a verb
- can significantly reduce the search space
- easier to list than annotating data (a few key rules)
- enforce rather than model (avoid non-local features)
- enable simple, painless NLP, e.g., joint inference

“Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

**Constraints:** Supervised and Unsupervised

- compact summaries of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly reduce the search space
  - easier to list than annotating data (a few key rules)
  - enforce rather than model (avoid non-local features)
  - enable simple, painless NLP, e.g., joint inference
    “Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

- relevant to unsupervised learning (less rope to hang self)
**Constraints: Supervised and Unsupervised**

- **compact summaries** of high-level insights into a domain
  — e.g., a sentence should have a verb
  — can significantly **reduce** the search space
  — easier to list than annotating data  (a few key rules)
  — enforce rather than model  (avoid non-local features)
  — enable simple, painless NLP, e.g., **joint inference**

  “Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

- relevant to **unsupervised learning**  (less rope to hang self)
  — inherently underconstrained problems...
Overview

**Constraints: Supervised and Unsupervised**

- **compact summaries** of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly **reduce** the search space
  - easier to list than annotating data (a few key rules)
  - enforce rather than model (avoid non-local features)
  - enable simple, painless NLP, e.g., **joint inference**
    “Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

- relevant to **unsupervised learning** (less rope to hang self)
  - inherently underconstrained problems...
  - in general, steer at the “right” regularities in data
**Constraints: Supervised and Unsupervised**

- **compact summaries** of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly **reduce** the search space
  - easier to list than annotating data  (a few key rules)
  - enforce rather than model  (avoid non-local features)
  - enable simple, painless NLP, e.g., **joint inference**
    “Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

- relevant to **unsupervised learning**  (less rope to hang self)
  - inherently underconstrained problems...
  - in general, steer at the “right” regularities in data
  - specifically, useful for **grammar (parser) induction**
Constraints: Supervised and Unsupervised

- compact summaries of high-level insights into a domain
  - e.g., a sentence should have a verb
  - can significantly reduce the search space
  - easier to list than annotating data (a few key rules)
  - enforce rather than model (avoid non-local features)
  - enable simple, painless NLP, e.g., joint inference
    “Integer Linear Programming in NLP” Tutorial (Chang et al., 2010)

- relevant to unsupervised learning (less rope to hang self)
  - inherently underconstrained problems...
  - in general, steer at the “right” regularities in data
  - specifically, useful for grammar (parser) induction
  - linguistic structure underdetermined by raw text
Constraints: Parser and Grammar Induction
Constraints: Parser and Grammar Induction

- the model
Constraints: Parser and Grammar Induction

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

Constraints: Parser and Grammar Induction

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

- partial *bracketings*  (Pereira and Schabes, 1992)
Constraints: Parser and Grammar Induction

- the model, e.g., projective trees (Klein and Manning, 2004) — Dependency Model with Valence (DMV)
- partial bracketings (Pereira and Schabes, 1992)
- synchronous grammars (Alshawi and Douglas, 2000)
Constraints: Parser and Grammar Induction

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

- partial **bracketings** (Pereira and Schabes, 1992)
- **synchronous** grammars (Alshawi and Douglas, 2000)
- **linear-time** parsing (Seginer, 2007)
Constraints: Parser and Grammar Induction

- the model, e.g., **projective** trees  \((\text{Klein and Manning, 2004})\)
  - Dependency Model with Valence (DMV)

- partial **bracketings**  \((\text{Pereira and Schabes, 1992})\)
- **synchronous** grammars  \((\text{Alshawi and Douglas, 2000})\)
- **linear-time** parsing  \((\text{Seginer, 2007})\)
- **skewness** of trees  \((\text{Seginer, 2007})\)
Constraints: Parser and Grammar Induction

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

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

- **synchronous grammars**  
  (Alshawi and Douglas, 2000)

- **linear-time** parsing  
  (Seginer, 2007)

- **skewness** of trees

- **Zipfian** distribution of words  
  (Seginer, 2007)
Constraints: Parser and Grammar Induction

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

- **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)
Constraints: Parser and Grammar Induction

- the model, e.g., \textbf{projective} trees \hfill (Klein and Manning, 2004)
  
  \textit{—} \textbf{Dependency Model with Valence (DMV)}

- \textbf{partial} \textit{bracketings} \hfill (Pereira and Schabes, 1992)
- \textbf{synchronous} grammars \hfill (Alshawi and Douglas, 2000)
- \textbf{linear-time} parsing \hfill (Seginer, 2007)
- \textbf{skewness} of trees \hfill (Seginer, 2007)
- \textbf{Zipfian} \textit{distribution} of \textit{words} \hfill (Seginer, 2007)
- \textbf{sparse} \textit{posterior} \textit{regularization} \hfill (Ganchev et al., 2009)
Constraints: Parser and Grammar Induction

- the model, e.g., 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)
Constraints: Partial Bracketings
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  — improved time complexity per iteration
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments

- problem: requires supervision (worst case — parse trees)

- how to make it work, in the absence of a treebank?
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments

- **problem**: requires supervision (worst case — parse trees)

how to make it work, in the *absence* of a treebank?
  - more, partially annotated corpora:
Constraints: Partial Bracketings

- play well with **EM** *(Pereira and Schabes, 1992)*
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments

- **problem**: requires supervision (worst case — parse trees)

- how to make it work, in the absence of a *treebank*?
  - more, partially annotated *corpora*:
    - English POS chunking *(Chen and Lee, 1995)*
Overview

Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)

how to make it work, in the absence of a treebank?

- more, partially annotated corpora:
  - English POS chunking (Chen and Lee, 1995)
  - Japanese clause splitting (Inui and Kotani, 2001)
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)

- how to make it work, in the absence of a treebank?
  - more, partially annotated corpora:
    - English POS chunking (Chen and Lee, 1995)
    - Japanese clause splitting (Inui and Kotani, 2001)
  - our approach:
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)

- how to make it work, in the absence of a treebank?
  - more, partially annotated corpora:
    - English POS chunking (Chen and Lee, 1995)
    - Japanese clause splitting (Inui and Kotani, 2001)
  - our approach:
    - would like to scale up to the web anyway
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)

how to make it work, in the absence of a treebank?

  - more, partially annotated corpora:
    - English POS chunking (Chen and Lee, 1995)
    - Japanese clause splitting (Inui and Kotani, 2001)
  - our approach:
    - would like to scale up to the web anyway
    - use what’s at hand (Verne, 1873; 1972)
Overview

Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments
- problem: requires supervision (worst case — parse trees)

how to make it work, in the absence of a treebank?

  ▶ more, partially annotated corpora:
    - English POS chunking (Chen and Lee, 1995)
    - Japanese clause splitting (Inui and Kotani, 2001)
  ▶ our approach:
    - would like to scale up to the web anyway
    - use what’s at hand (Verne, 1873; 1972)
    - HTML structure
Constraints: Partial Bracketings

- play well with EM (Pereira and Schabes, 1992)
  - improved time complexity per iteration
  - fewer iterations to reach a good grammar
  - better agreement with qualitative judgments

- problem: requires supervision (worst case — parse trees)

- how to make it work, in the absence of a treebank?
  - more, partially annotated corpora:
    - English POS chunking (Chen and Lee, 1995)
    - Japanese clause splitting (Inui and Kotani, 2001)
  - our approach:
    - would like to scale up to the web anyway
    - use what’s at hand (Verne, 1873; 1972)
    - HTML structure

- solution: mark-up!
Web Mark-Up: Diamonds in the Rough

suggestive example:

..., whereas McCain is secure on the topic, Obama <a>[VP worries about winning the pro-Israel vote]</a>.
Web Mark-Up: Diamonds in the Rough

suggestive example:

..., whereas McCain is secure on the topic, Obama <a>[VP worries about winning the pro-Israel vote]</a>. Intuition: diamonds in the rough.
Web Mark-Up: Diamonds in the Rough

suggestive example:

..., whereas McCain is secure on the topic, Obama \(<a>[v_P \text{ worries about winning the pro-Israel vote}]</a>\).

intuition:
diamonds in the rough

- natural language pre-processing (NLPP?):
Web Mark-Up: Diamonds in the Rough

suggestive example:

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

intuition: diamonds in the rough

natural language pre-processing (NLPP?): — stripping out HTML is an ugly chore...
Web Mark-Up: Diamonds in the Rough

suggestive example:

..., whereas McCain is secure on the topic, Obama \(<a>[v_p \text{ worries about winning the pro-Israel vote}]</a>\).

intuition:

diamonds
in the
rough

natural language pre-processing (NLPP?):
— stripping out HTML is an ugly chore...
— instead of rushing to discard it, try polishing!
Outline:

- structure of this talk (as guided by the time constraints):
Outline:

- structure of this talk (as guided by the time constraints):
  - linguistic analysis of a single blog
Outline:

- structure of this talk (as guided by the time constraints):
  - linguistic analysis of a single blog
    — is there a connection between syntax and mark-up?
Outline:

- structure of this talk (as guided by the time constraints):
  1. **linguistic analysis** of a single blog
     - is there a connection between *syntax* and *mark-up*?
     - yes...
Outline:

- structure of this talk (as guided by the time constraints):
  - **linguistic analysis** of a single blog
    - is there a connection between *syntax* and *mark-up*?
    - yes... (but what is it? and is it useful?)
Outline:

- structure of this talk (as guided by the time constraints):
  1. linguistic analysis of a single blog
     — is there a connection between syntax and mark-up?
     — yes... (but what is it? and is it useful?)
  2. proposed parsing constraints, refined from mark-up
Outline:

- structure of this talk (as guided by the time constraints):
  1. linguistic analysis of a single blog
     — is there a connection between syntax and mark-up?
     — yes... (but what is it? and is it useful?)
  2. proposed parsing constraints, refined from mark-up
  3. experimental results for unsupervised dependency parsing
Outline:

• structure of this talk (as guided by the time constraints):

1. linguistic analysis of a single blog
   — is there a connection between syntax and mark-up?
   — yes... (but what is it? and is it useful?)

2. proposed parsing constraints, refined from mark-up

3. experimental results for unsupervised dependency parsing
   — parsed the web
Outline:

- structure of this talk (as guided by the time constraints):
  1. **linguistic analysis** of a single blog
     - is there a connection between **syntax** and **mark-up**?
     - yes... (but what is it? and is it useful?)
  2. proposed **parsing constraints**, refined from mark-up
  3. **experimental results** for unsupervised dependency parsing
     - parsed the web
     - but you don’t have to...
Outline:

- structure of this talk (as guided by the time constraints):
  1. linguistic analysis of a single blog
     - is there a connection between syntax and mark-up?
     - yes... (but what is it? and is it useful?)
  2. proposed parsing constraints, refined from mark-up
  3. experimental results for unsupervised dependency parsing
     - parsed the web
     - but you don’t have to...
     - best results with just the blog
Outline:

- structure of this talk (as guided by the time constraints):
  1. linguistic analysis of a single blog
     - is there a connection between syntax and mark-up?
     - yes... (but what is it? and is it useful?)
  2. proposed parsing constraints, refined from mark-up
  3. experimental results for unsupervised dependency parsing
     - parsed the web
     - but you don’t have to...
     - best results with just the blog
     - ... web news also state-of-the-art
Outline:

- structure of this talk (as guided by the time constraints):
  1. **linguistic analysis** of a single blog
     - is there a connection between *syntax* and *mark-up*?
     - yes... (but what is it? and is it useful?)
  2. proposed **parsing constraints**, refined from mark-up
  3. **experimental results** for unsupervised dependency parsing
     - parsed the web
     - but you don’t have to...
     - best results with just the blog
     - ... web news also state-of-the-art

- minor yet recurring theme: **less is more**
Outline:

- dropped details:
Outline:

- dropped details:
  - model: Dependency Model with Valence (DMV) [POS tags] (Klein and Manning, 2004)
Outline:

- **dropped details:**
  - **model:** Dependency Model with Valence (DMV) [POS tags] (Klein and Manning, 2004)
  - **learning engine:** Viterbi EM (not Inside-Outside) [CoNLL] (Spitkovsky et al., 2010)
Outline:

- **dropped details:**
  - **model:** Dependency Model with Valence (DMV)  
    [POS tags] (Klein and Manning, 2004)
  - **learning engine:**  Viterbi EM (not Inside-Outside)  
    [CoNLL] (Spitkovsky et al., 2010)
  - **methodology:** experimental design (hundreds of runs)  
    [ACL] (Spitkovsky et al., 2010)
Data:
Data:

- variety of data-set sizes and genres:
Data:

- variety of data-set sizes and genres:
  — from biggest to smallest
Data:

- variety of data-set *sizes* and *genres*:
  — from biggest to smallest, from messiest to cleanest
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

1. English web
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest
    - English web
      - nearly $100B$ POS tokens

http://google.com/en
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

  1. English web
     - nearly $100B$ POS tokens
     - TnT-tagged (Brants, 2000)
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest
  1. English web
     - nearly $100B$ POS tokens
     - TnT-tagged (Brants, 2000)
  2. web news
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

  1. English web
     - nearly 100B POS tokens
     - TnT-tagged

  2. web news
     - about 30B tokens

http://google.com/en
(Brants, 2000)
http://news.google.com/
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

1. English web
   - nearly $100B$ POS tokens
   - TnT-tagged

2. web news
   - about $30B$ tokens

3. political opinion blog
Data:

- **variety of data-set sizes and genres:**
  - from biggest to smallest, from messiest to cleanest

  1. **English web**
     - nearly $100B$ POS tokens
     - TnT-tagged
     
  2. **web news**
     - about $30B$ tokens
     
  3. **political opinion blog**
     - a little over $1M$ tokens
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

1. English web
   - nearly 100B POS tokens
   - TnT-tagged (Brants, 2000)

2. web news
   - about 30B tokens

3. political opinion blog
   - a little over 1M tokens
   - manually cleaned up (for analysis)
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest
  1. English web
     - nearly 100B POS tokens
     - TnT-tagged (Brants, 2000)
  2. web news
     - about 30B tokens
  3. political opinion blog
     - a little over 1M tokens
     - manually cleaned up (for analysis)
     - Charniak-parsed (Charniak and Johnson, 2005)
Data:

- **variety of data-set sizes and genres:**
  - from biggest to smallest, from messiest to cleanest
  1. **English web**
     - nearly 100B POS tokens
     - TnT-tagged
  2. **web news**
     - about 30B tokens
  3. **political opinion blog**
     - a little over 1M tokens
     - manually cleaned up (for analysis)
       - Charniak-parsed
       - Stanford-tagged

http://news.google.com/ (Toutanova et al., 2003)
Overview

Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

1. **English web**
   - nearly $100B$ POS tokens
   - TnT-tagged (Brants, 2000)

2. **web news**
   - about $30B$ tokens

3. **political opinion blog**
   - a little over $1M$ tokens
   - manually cleaned up (for analysis)
   - Charniak-parsed (Charniak and Johnson, 2005)
   - Stanford-tagged (Toutanova et al., 2003)

4. **WSJ** — just over $1M$ tokens (Marcus et al., 1993)
Data:

- variety of data-set sizes and genres:
  - from biggest to smallest, from messiest to cleanest

1. English web
   - nearly $100B$ POS tokens
   - TnT-tagged

2. web news
   - about $30B$ tokens

3. political opinion blog
   - a little over $1M$ tokens
   - manually cleaned up (for analysis)
     * Charniak-parsed
     * Stanford-tagged

4. WSJ — just over $1M$ tokens

5. Brown — under $400K$ tokens
### Syntax of Mark-Up: POS Sequences $\langle a, b, i, u \rangle$

<table>
<thead>
<tr>
<th>Sequence</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP NNP</td>
<td>16.1</td>
</tr>
<tr>
<td>NNP</td>
<td>8.3</td>
</tr>
<tr>
<td>NNP NNP NNP</td>
<td>5.4</td>
</tr>
<tr>
<td>NN</td>
<td>5.4</td>
</tr>
<tr>
<td>JJ NN</td>
<td>2.6</td>
</tr>
<tr>
<td>DT NNP NNP</td>
<td>1.8</td>
</tr>
<tr>
<td>NNS</td>
<td>1.8</td>
</tr>
<tr>
<td>JJ</td>
<td>1.5</td>
</tr>
<tr>
<td>VBD</td>
<td>1.3</td>
</tr>
<tr>
<td>DT NNP NNP NNP</td>
<td>1.2</td>
</tr>
<tr>
<td>JJ NNS</td>
<td>1.1</td>
</tr>
<tr>
<td>NNP NN</td>
<td>1.0</td>
</tr>
<tr>
<td>NN NN</td>
<td>1.0</td>
</tr>
<tr>
<td>VBN</td>
<td>0.8</td>
</tr>
<tr>
<td>NNP NNP NNP NNP</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50.0</strong></td>
</tr>
</tbody>
</table>
### Syntax of Mark-Up: Dominating Non-Terminals

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>74.5</td>
</tr>
<tr>
<td>VP</td>
<td>12.9</td>
</tr>
<tr>
<td>S</td>
<td>6.8</td>
</tr>
<tr>
<td>PP</td>
<td>1.6</td>
</tr>
<tr>
<td>ADJP</td>
<td>0.9</td>
</tr>
<tr>
<td>FRAG</td>
<td>0.8</td>
</tr>
<tr>
<td>ADVP</td>
<td>0.5</td>
</tr>
<tr>
<td>SBAR</td>
<td>0.5</td>
</tr>
<tr>
<td>PRN</td>
<td>0.2</td>
</tr>
<tr>
<td>NX</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>99.0</strong></td>
</tr>
</tbody>
</table>
Syntax of Mark-Up: Common Constituents

..., but \[S [NP the <a>Toronto Star</a>] [VP reports [NP this] [PP in the softest possible way]]</a>, [S stating ...]]]
Syntax of Mark-Up: Common Constituents

..., but \[ S \rightarrow NP \_VP \]

\[ S \rightarrow [NP \_Toronto \ Star][VP \ reports \ [NP \_this][PP \ in \ the \ softest \ possible \ way]][S \ stating \ ...]] \]

---

Spitkovsky et al. (Stanford & Google) Profiting from Mark-Up ACL (2010-07-14) 11 / 33
Syntax of Mark-Up: Common Constituents

..., but \[ S \rightarrow [NP \text{ the } \langle a \rangle \text{Toronto Star}] [VP \text{ reports } [NP \text{ this}] [PP \text{ in the softest possible way}] \langle /a \rangle, [S \text{ stating } ...]] ] \]

\[ S \rightarrow NP \_VP \rightarrow DT \_NNP \_NNP \_VBZ \_NP \_PP \_S \]
Syntax of Mark-Up: Constituent Productions

<table>
<thead>
<tr>
<th>Production</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP → NNP NNP</td>
<td>9.6</td>
</tr>
<tr>
<td>NP → NNP</td>
<td>4.6</td>
</tr>
<tr>
<td>NP → NP PP</td>
<td>3.4</td>
</tr>
<tr>
<td>NP → NNP NNP NNP</td>
<td>2.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>2.1</td>
</tr>
<tr>
<td>NP → NN</td>
<td>1.8</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NN</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>1.6</td>
</tr>
<tr>
<td>S → NP VP</td>
<td>1.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.2</td>
</tr>
<tr>
<td>NP → DT JJ NN</td>
<td>1.1</td>
</tr>
<tr>
<td>NP → NNS</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → JJ NN</td>
<td>0.8</td>
</tr>
<tr>
<td>NP → NP NP</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>35.3</td>
</tr>
</tbody>
</table>
# Syntax of Mark-Up: Constituent Productions

<table>
<thead>
<tr>
<th>Production</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP → NNP</td>
<td>9.6</td>
</tr>
<tr>
<td>NP → NNP NNP</td>
<td>4.6</td>
</tr>
<tr>
<td>NP → NP PP</td>
<td>3.4</td>
</tr>
<tr>
<td>NP → NNP NNP NNP</td>
<td>2.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>2.1</td>
</tr>
<tr>
<td>NP → NN</td>
<td>1.8</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NN</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>1.6</td>
</tr>
<tr>
<td>S → NP VP</td>
<td>1.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.2</td>
</tr>
<tr>
<td>NP → DT JJ NN</td>
<td>1.1</td>
</tr>
<tr>
<td>NP → NNS</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → JJ NN</td>
<td>0.8</td>
</tr>
<tr>
<td>NP → NP NP</td>
<td>0.8</td>
</tr>
</tbody>
</table>

35.3
## Syntax of Mark-Up: Constituent Productions

<table>
<thead>
<tr>
<th>Production</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP → NNP NNP</td>
<td>9.6</td>
</tr>
<tr>
<td>NP → NNP</td>
<td>4.6</td>
</tr>
<tr>
<td>NP → NP PP</td>
<td>3.4</td>
</tr>
<tr>
<td>NP → NNP NNP NNP</td>
<td>2.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>2.1</td>
</tr>
<tr>
<td>NP → NN</td>
<td>1.8</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NN</td>
<td>1.7</td>
</tr>
<tr>
<td>NP → DT NNP NNP</td>
<td>1.6</td>
</tr>
<tr>
<td>S → NP VP</td>
<td>1.4</td>
</tr>
<tr>
<td>NP → DT NNP NNP NNP</td>
<td>1.2</td>
</tr>
<tr>
<td>NP → DT JJ NN</td>
<td>1.1</td>
</tr>
<tr>
<td>NP → NNS</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → JJ NN</td>
<td>0.8</td>
</tr>
<tr>
<td>NP → NP NP</td>
<td>0.8</td>
</tr>
</tbody>
</table>

35.3
Syntax of Mark-Up: Common Dependencies

..., but $[S [NP the <a>Toronto Star</a>] [VP reports [NP this] [PP in the softest possible way]]$, stating ...
Syntax of Mark-Up: Common Dependencies

..., but [S [NP the <a>Toronto Star] [VP reports [NP this] [PP in the softest possible way]<a>], [S stating ...]]]
Syntax of Mark-Up: Common Dependencies

..., but \[ S [\text{NP the } \texttt{Toronto Star}] [\text{VP reports } \text{NP this}] [\text{PP in the softest possible way}] ] , [S stating ...]]

\[ DT \text{ NNP NNP VBZ DT IN DT JJS JJ NN} \]

\[ DT \text{ NNP VBZ} \]
Syntax of Mark-Up: Common Dependencies

..., but [S [NP the <a>Toronto Star</a>] [VP reports [NP this] [PP in the softest possible way]]</a>, [S stating ...]]

DT NNP NNP VBZ DT IN DT JJS JJ NN

DT NNP VBZ

“the <a>Star reports</a>”
Syntax of Mark-Up: Head-Outward Spawns

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>24.4</td>
</tr>
<tr>
<td>NN</td>
<td>8.1</td>
</tr>
<tr>
<td>DT</td>
<td>6.1</td>
</tr>
<tr>
<td>NNP</td>
<td>5.9</td>
</tr>
<tr>
<td>NNS</td>
<td>4.5</td>
</tr>
<tr>
<td>NNPS</td>
<td>1.4</td>
</tr>
<tr>
<td>VBG</td>
<td>1.3</td>
</tr>
<tr>
<td>NNP</td>
<td>1.2</td>
</tr>
<tr>
<td>VBD</td>
<td>1.0</td>
</tr>
<tr>
<td>IN</td>
<td>1.0</td>
</tr>
<tr>
<td>VBN</td>
<td>1.0</td>
</tr>
<tr>
<td>DT</td>
<td>0.9</td>
</tr>
<tr>
<td>JJ</td>
<td>0.9</td>
</tr>
<tr>
<td>VBZ</td>
<td>0.9</td>
</tr>
<tr>
<td>POS</td>
<td>0.9</td>
</tr>
<tr>
<td>JJ</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Total: 59.4%

Spitkovsky et al. (Stanford & Google) Profiting from Mark-Up ACL (2010-07-14)
Syntax of Mark-Up: Head-Outward Spawns

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>24.4</td>
</tr>
<tr>
<td>NN</td>
<td>8.1</td>
</tr>
<tr>
<td>DT NNP</td>
<td>6.1</td>
</tr>
<tr>
<td>DT NN</td>
<td>5.9</td>
</tr>
<tr>
<td>NNS</td>
<td>4.5</td>
</tr>
<tr>
<td>NNPS</td>
<td>1.4</td>
</tr>
<tr>
<td>VBG</td>
<td>1.3</td>
</tr>
<tr>
<td>NNP NNP NN</td>
<td>1.2</td>
</tr>
<tr>
<td>VBD</td>
<td>1.0</td>
</tr>
<tr>
<td>IN</td>
<td>1.0</td>
</tr>
<tr>
<td>VBN</td>
<td>1.0</td>
</tr>
<tr>
<td>DT JJ NN</td>
<td>0.9</td>
</tr>
<tr>
<td>VBZ</td>
<td>0.9</td>
</tr>
<tr>
<td>POS NNP</td>
<td>0.9</td>
</tr>
<tr>
<td>JJ</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>59.4</strong></td>
</tr>
</tbody>
</table>
## Syntax of Mark-Up: Head-Outward Spawns

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP</td>
<td>24.4</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>NNS</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>NNPS</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>VBG</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>NNP NNP NN</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>VBD</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>VBN</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>DT JJ NN</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>VBZ</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>POS NNP</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>59.4</td>
<td></td>
</tr>
</tbody>
</table>
Syntax of Mark-Up: Exception

... [NP a 1994 <i>New Yorker</i> article] ...

Spitkovsky et al. (Stanford & Google)  Profiting from Mark-Up  ACL (2010-07-14)
Syntax of Mark-Up: Exception

... [NP a 1994 <i>New Yorker</i> article] ...

- consequence of bare NPs
  — ... and “head percolation” rules
Syntax of Mark-Up: Summary

- not just single words
- (lots of long noun phrases)
Syntax of Mark-Up: Summary

- not just single words (lots of long noun phrases)
- some verbs, adjectives, etc. (i.e., not just nouns)
Syntax of Mark-Up: Summary

- not just single words (lots of long noun phrases)
- some verbs, adjectives, etc. (i.e., not just nouns)
- apparent agreement with constituents
**Syntax of Mark-Up: Summary**

- not just single words (lots of long noun phrases)
- some verbs, adjectives, etc. (i.e., not just nouns)
- apparent agreement with constituents
- and also with dependencies
**Syntax of Mark-Up: Summary**

- not just single words (lots of long noun phrases)
- some verbs, adjectives, etc. (i.e., not just nouns)

- apparent agreement with constituents

- and also with dependencies

— but is there enough mark-up?
Syntax of Mark-Up: Summary

- not just single words  
  (lots of long noun phrases) 
- some verbs, adjectives, etc.  
  (i.e., not just nouns) 
- apparent agreement with constituents 
- and also with dependencies

— but is there enough mark-up?

- 11% of all sentences in the blog are annotated
Syntax of Mark-Up: Summary

- not just single words (lots of long noun phrases)
- some verbs, adjectives, etc. (i.e., not just nouns)
- apparent agreement with constituents
- and also with dependencies

— but is there enough mark-up?

- 11% of all sentences in the blog are annotated
- 9% have multi-token bracketings
Proposed Constraints: Constituents?
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in [NP an analysis] PP of perhaps the most astonishing PC item I have yet stumbled upon].
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in \([\text{NP} <a> \text{NP an analysis} </a> \text{PP} \text{of perhaps the most astonishing PC item I have yet stumbled upon}]\).

- these are rough diamonds...
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in \[ \text{NP}<a>\text{NP an analysis}</a>[\text{PP of perhaps the most astonishing PC item I have yet stumbled upon}]. \]

- these are rough diamonds...

- many disagreements due to treebank idiosyncrasies:
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in \([\text{NP} < a > \text{NP an analysis}] < / a > \text{PP of perhaps the most astonishing PC item I have yet stumbled upon}]\).

- these are rough diamonds...

- many disagreements due to treebank idiosyncrasies:
  — bare NPs (internal structure)
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in \[ NP \text{an analysis PP} \text{of perhaps the most astonishing PC item I have yet stumbled upon}]\].

- these are rough diamonds...

- many disagreements due to treebank idiosyncrasies:
  - bare NPs (internal structure)
  - N-bars (missing determiners)
Proposed Constraints: Constituents?

- 48.0% agreement with Charniak’s trees, e.g.,

  ... in $[\text{NP< }a\text{>}[\text{NP an analysis }]<a>\text{[PP of perhaps the most astonishing PC item I have yet stumbled upon}]]$.

- these are rough diamonds...

- many disagreements due to treebank idiosyncrasies:
  - bare NPs (internal structure)
  - N-bars (missing determiners)

- ... but we’ll polish them anyway!
Proposed Constraints: Dependencies!
Proposed Constraints: Dependencies!

- a more stylistically-forgiving framework
Proposed Constraints: Dependencies!

- a more stylistically-forgiving framework
- start with the strictest possible constraint
Proposed Constraints: Dependencies!

- a more stylistically-forgiving framework
- start with the strictest possible constraint
- then slowly relax it
Proposed Constraints: Dependencies!

- a more stylistically-forgiving framework
- start with the strictest possible constraint
- then slowly relax it
- every example demonstrating a softer constraint doubles as a counter-example against all previous
Proposed Constraints: Strict
Proposed Constraints: Strict

- seal mark-up into attachments
Proposed Constraints: Strict

- seal mark-up into attachments, e.g.,

  As author of `<i>The Satanic Verses</i>`, I ...
Proposed Constraints: Strict

- seal mark-up into attachments, e.g.,

As author of *The Satanic Verses*, I ...

— just 35.6% agreement with head-percolated trees
Proposed Constraints: Loose
Proposed Constraints: Loose

- allow bracketed head word external dependents
Proposed Constraints: Loose

- allow bracketed **head** word external dependents, e.g.,

  ... the *Toronto Star* reports ...

[Diagram showing the structure of the sentence with the 'Toronto Star' highlighted and annotated with the tag `<i>`]
Proposed Constraints: **Loose**

- allow bracketed *head* word external dependents, e.g.,

  ... the `<i>Toronto Star</i>` reports ...

  — already 87.5% agreement with head-percolated trees
Proposed Constraints: Sprawl
Proposed Constraints: Sprawl

- allow all bracketed words external dependents
Proposed Constraints: Sprawl

- allow all bracketed words external dependents, e.g.,

... the `<a>`Toronto Star reports` ... `<a>` ...
Proposed Constraints: Sprawl

- allow all bracketed words external dependents, e.g.,

  ... the `<a>` Toronto Star reports ... `</a>` ...

— now 95.1% agreement with head-percolated trees
Proposed Constraints: Tear
Proposed Constraints: Tear

- fracture by same-side external heads
Proposed Constraints: Tear

- fracture by same-side external heads, e.g.,

... concession ... has raised eyebrows among those waiting for Fox News in Canada.
Proposed Constraints: Tear

- fracture by same-side external heads, e.g.,

... concession ... has raised eyebrows among those waiting for Fox News in Canada.

— finally, 98.9% agreement with head-percolated trees
Proposed Constraints: Summary
Proposed Constraints: Summary

- remaining 1.1% mostly due to parser errors...
Proposed Constraints: Summary

- remaining 1.1% mostly due to parser errors...
  ... found one (very rare) true negative disagreement
Proposed Constraints: Summary

- remaining 1.1% mostly due to parser errors...
  ... found one (very rare) true negative disagreement

- a suite of highly (88%, 95%, 99%) accurate constraints
Proposed Constraints: Summary

- remaining 1.1% mostly due to parser errors...
  ... found one (very rare) true negative disagreement

- a suite of highly (88%, 95%, 99%) accurate constraints,
  ... of varying degrees of informativeness
Proposed Constraints: Summary

- remaining 1.1% mostly due to parser errors...
  ... found one (very rare) true negative disagreement

- a suite of highly (88%, 95%, 99%) accurate constraints,
  ... of varying degrees of informativeness

- first two can easily guide Viterbi training!
## Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th>Incarnation</th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cohen and Smith, 2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Spitkovsky et al., 2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Headden et al., 2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th>Incarnation</th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cohen and Smith, 2009)</td>
<td>62.0</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>(Spitkovsky et al., 2010)</td>
<td>57.1</td>
<td>45.0</td>
<td>43.6</td>
</tr>
<tr>
<td>(Headden et al., 2009)</td>
<td>68.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th>Incarnation</th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cohen and Smith, 2009)</td>
<td>62.0</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>(Spitkovsky et al., 2010)</td>
<td>57.1</td>
<td>45.0</td>
<td>43.6</td>
</tr>
<tr>
<td>(Headden et al., 2009)</td>
<td>68.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
</tbody>
</table>

- **state-of-the-art** results
## Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th>Incarnation</th>
<th>WSJ10</th>
<th>WSJ$^\infty$</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cohen and Smith, 2009)</td>
<td>62.0</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>(Spitkovsky et al., 2010)</td>
<td>57.1</td>
<td>45.0</td>
<td>43.6</td>
</tr>
<tr>
<td>(Headden et al., 2009)</td>
<td>68.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td>+0.5</td>
<td>+5.4</td>
<td>+9.7</td>
</tr>
</tbody>
</table>

- **state-of-the-art** results

- **linguistic constraints** help with the task!
## Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WEB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Spitkovsky et al. (Stanford & Google) Profiting from Mark-Up ACL (2010-07-14) 25 / 33
**Experimental Results: Dependency Accuracy (%)**

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ(\infty)</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td>67.3</td>
<td>50.1</td>
<td>51.6</td>
</tr>
<tr>
<td>WEB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- no need to manually clean data!
Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td>67.3</td>
<td>50.1</td>
<td>51.6</td>
</tr>
<tr>
<td>WEB</td>
<td>64.1</td>
<td>46.3</td>
<td>46.9</td>
</tr>
</tbody>
</table>

- no need to manually clean data!
- nevertheless, less is more...
**Experimental Results: Dependency Accuracy (%)**

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td>67.3</td>
<td>50.1</td>
<td>51.6</td>
</tr>
<tr>
<td>WEB</td>
<td>64.1</td>
<td>46.3</td>
<td>46.9</td>
</tr>
</tbody>
</table>

- no need to manually clean data!
- nevertheless, less is more...
- loose constraint consistently delivers best results
Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td>67.3</td>
<td>50.1</td>
<td>51.6</td>
</tr>
<tr>
<td>WEB</td>
<td>64.1</td>
<td>46.3</td>
<td>46.9</td>
</tr>
</tbody>
</table>

- no need to manually clean data!
- nevertheless, less is more...
- loose constraint consistently delivers best results
- requires domain adaptation (re-training on WSJ)
Experimental Results: Dependency Accuracy (%)

<table>
<thead>
<tr>
<th></th>
<th>WSJ10</th>
<th>WSJ∞</th>
<th>Brown100</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOG</td>
<td>69.3</td>
<td>50.4</td>
<td>53.3</td>
</tr>
<tr>
<td>NEWS</td>
<td>67.3</td>
<td>50.1</td>
<td>51.6</td>
</tr>
<tr>
<td>WEB</td>
<td>64.1</td>
<td>46.3</td>
<td>46.9</td>
</tr>
</tbody>
</table>

- no need to manually clean data!
- nevertheless, less is more...
- loose constraint consistently delivers best results
- requires domain adaptation (re-training on WSJ)
- perhaps bigger gains if lexicalized?
Experimental Results: Why Didn’t the Web Help?
Experimental Results: Why Didn’t the Web Help?

- language identification
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
- boiler-plate
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
- boiler-plate, POS-tagging:
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
- boiler-plate, POS-tagging:

<table>
<thead>
<tr>
<th>POS Sequence</th>
<th>WEB Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample web sentence, chosen uniformly at random.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>DT NNS VBN</td>
</tr>
<tr>
<td>2</td>
<td>NNP NNP NNP</td>
</tr>
<tr>
<td>3</td>
<td>NN IN TO VB RB</td>
</tr>
<tr>
<td>4</td>
<td>NN IN IN PRP$ JJ NN</td>
</tr>
</tbody>
</table>

Yuasa et al.

Sign in to YouTube now!

Sign in with your Google Account!
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
- boiler-plate, POS-tagging:

<table>
<thead>
<tr>
<th>POS Sequence</th>
<th>WEB Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample web sentence, chosen uniformly at random.</td>
<td>All rights reserved.</td>
</tr>
<tr>
<td>1 DT NNS VBN</td>
<td>82,858,487</td>
</tr>
<tr>
<td>2 NNP NNP NNP</td>
<td>65,889,181</td>
</tr>
<tr>
<td>3 NN IN TO VB RB</td>
<td>31,007,783</td>
</tr>
<tr>
<td>4 NN IN IN PRP$ JJ NN</td>
<td>31,007,471</td>
</tr>
</tbody>
</table>
Experimental Results: Why Didn’t the Web Help?

- language identification, sentence-breaking
- boiler-plate, POS-tagging:

<table>
<thead>
<tr>
<th>POS Sequence</th>
<th>WEB Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample web sentence, chosen uniformly at random.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>DT NNS VBN</td>
</tr>
<tr>
<td>2</td>
<td>NNP NNP NNP</td>
</tr>
<tr>
<td>3</td>
<td>NN IN TO VB RB</td>
</tr>
<tr>
<td>4</td>
<td>NN IN IN PRP$ JJ NN</td>
</tr>
</tbody>
</table>

All rights reserved. Yuasa et al.

Sign in to YouTube now!

Sign in with your Google Account!

- ambiguous noun phrases: “click here” and “print post”
Summary
Summary

- strong connection between mark-up and syntax


**Summary**

- strong **connection** between mark-up and syntax
- state-of-the-art **unsupervised dependency** parsing
Summary

- strong **connection** between mark-up and syntax
- state-of-the-art **unsupervised dependency** parsing
- other **parsing** applications:
Summary

- strong connection between mark-up and syntax
- state-of-the-art unsupervised dependency parsing
- other parsing applications:
  - supervised parsing (via self-training)
Summary

- strong connection between mark-up and syntax
- state-of-the-art unsupervised dependency parsing
- other parsing applications:
  - supervised parsing (via self-training)
  - constituent parsing (via discriminative features)
Summary

- strong connection between mark-up and syntax
- state-of-the-art unsupervised dependency parsing
- other parsing applications:
  - supervised parsing (via self-training)
  - constituent parsing (via discriminative features)
  - balanced punctuation? (e.g., quotes and parens)
Potential
Potential

- another motivating example:
Potential

- another motivating example:

\[
[NP \ [NP \ \text{Libyan ruler}] \ \text{<a>}[NP \ \text{Mu‘ammar al-Qaddafi}]</a>]
\]
Potential

• another motivating example:

\[
\text{[NP \ [NP Libyan ruler] <a>[NP Mu‘ammar al-Qaddafi]</a> ]}
\]

— internal structure of a compound

(Vadas and Curran, 2007)
Potential

- another motivating **example**:

  $$\text{[NP [NP Libyan ruler] <a>[NP Mu‘ammar al-Qaddafi</a>]}$$

  — **internal structure** of a compound  
    (Vadas and Curran, 2007)

  — **lower-level** tokenization signal
Potential

- another motivating example:

\[
\text{[NP [NP Libyan ruler] <a>[NP Mu‘ammar al-Qaddafi]</a>]
\]

— internal structure of a compound

(Vadas and Curran, 2007)

— lower-level tokenization signal

http://nlp.stanford.edu:8080/parser/

\[
\text{(NP (ADJP (NP (JJ Libyan) (NN ruler)))}
\text{(JJ Mu))}
\text{("‘") (NN ammar) (NNS al-Qaddafi))}
\]
Potential

- other structured tasks in NLP:
Potential

- other structured tasks in NLP:
  - NE-tagging
Potential

- other structured tasks in NLP:
  - NE-tagging
  - NP-chunking
Potential

other structured tasks in NLP:

- NE-tagging
- NP-chunking
- CJK-segmentation
Potential

- other structured tasks in NLP:
  - NE-tagging
  - NP-chunking
  - CJK-segmentation
  - sentence-breaking
Potential

- other structured tasks in NLP:
  - NE-tagging
  - NP-chunking
  - CJK-segmentation
  - sentence-breaking

... and so forth!
Open Questions:
Open Questions:

- does this generalize to other genres?
Open Questions:

- does this generalize to other genres?
- does this generalize to other languages?
Open Questions:

- does this generalize to other genres?
- does this generalize to other languages?
- what would be the impact of lexicalization?
**Open Questions:**

- does this generalize to other **genres**?
- does this generalize to other **languages**?
- what would be the impact of **lexicalization**?
- are there broader **NLP implications**?
Conclusion

What We Make Available:
What We Make Available:

- all of our cleaned up annotations of the blog
What We Make Available:

- all of our cleaned up annotations of the blog
- a complete analysis of every annotated sentence
What We Make Available:

- all of our cleaned up annotations of the blog
- a complete analysis of every annotated sentence
- and the best (blog) model
What We Make Available:

- all of our cleaned up annotations of the blog

- a complete analysis of every annotated sentence

- and the best (blog) model

http://cs.stanford.edu/~valentin/
Thanks!

Questions?
Proposed Constraints: Exception
Proposed Constraints: Exception

- remaining 1.1% mostly due to parser errors...
Proposed Constraints: Exception

- remaining 1.1% mostly due to parser errors...

- a (very rare) true negative disagreement:
Proposed Constraints: Exception

- remaining 1.1% mostly due to parser errors...
- a (very rare) true negative disagreement:

The French broadcasting authority, [CSA](http://example.com), banned ... Al-Manar satellite television from ...