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)

`http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

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)
  `http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

- 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)
  `http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

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

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)
  `http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

- 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)
  `http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

- 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)
  `http://l2r.cs.uiuc.edu/~danr/Talks/ILP-CCM-Tutorial-NAACL10.pdf`

- 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)**

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 (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)**

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)**

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)

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., 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., 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)

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)

Phileas Fogg

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**

Phileas Fogg

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!

Phileas Fogg

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>[$_{VP}$ 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>[$_{VP}$ 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...
  — 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):**

  1. **linguistic analysis of a single 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?

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):

  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?)

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

  1. **English web** `http://google.com/en`
     — nearly $100B$ POS tokens

Data:

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

  1. **English web** **http://google.com/en**
     — 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** `http://google.com/en`
     — 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** http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. **web news** http://news.google.com/
     — about $30B$ tokens

Data:

- **variety of data-set** **sizes** **and** **genres**:
  — from biggest to smallest, from messiest to cleanest
  1. **English** **web** `http://google.com/en`
     — nearly $100B$ **POS tokens**
     — TnT-tagged **(Brants, 2000)**
  2. **web** **news** `http://news.google.com/`
     — 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** http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. web **news** http://news.google.com/
     — about $30B$ tokens
  3. political opinion **blog** http://danielpipes.org/
     — a little over $1M$ tokens

Data:

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

1. **English web** **http://google.com/en**
   — nearly $100B$ POS tokens
   — TnT-tagged **(Brants, 2000)**
2. **web news** **http://news.google.com/**
   — about $30B$ tokens
3. **political opinion blog** **http://danielpipes.org/**
   — 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** http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. web **news** http://news.google.com/
     — about $30B$ tokens
  3. political opinion **blog** http://danielpipes.org/
     — 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** http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. web **news** http://news.google.com/
     — about $30B$ tokens
  3. political opinion **blog** http://danielpipes.org/
     — a little over $1M$ tokens
     — manually cleaned up (for analysis)
     - ⋆ Charniak-parsed (Charniak and Johnson, 2005)
     - ⋆ Stanford-tagged (Toutanova et al., 2003)

Data:

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

  1. English web http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. web news http://news.google.com/
     — about $30B$ tokens
  3. political opinion blog http://danielpipes.org/
     — 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** http://google.com/en
     — nearly $100B$ POS tokens
     — TnT-tagged (Brants, 2000)
  2. **web news** http://news.google.com/
     — about $30B$ tokens
  3. **political opinion blog** http://danielpipes.org/
     — 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)
  5. **Brown — under $400K$ tokens** (Francis and Kucera, 1979)

Syntax of Mark-Up: POS Sequences <a, b, i, u>

	%
NNP NNP	16.1
NNP	8.3
NNP NNP NNP	5.4
NN	5.4
JJ NN	2.6
DT NNP NNP	1.8
NNS	1.8
JJ	1.5
VBD	1.3
DT NNP NNP NNP	1.2
JJ NNS	1.1
NNP NN	1.0
NN NN	1.0
VBN	0.8
NNP NNP NNP NNP	0.8
	50.0

Syntax of Mark-Up: Dominating Non-Terminals

	%
NP	74.5
VP	12.9
S	6.8
PP	1.6
ADJP	0.9
FRAG	0.8
ADVP	0.5
SBAR	0.5
PRN	0.2
NX	0.2
	99.0

Syntax of Mark-Up: Common Constituents

…, but $[_{\text{S}}$ $[_{\text{NP}}$ the <a>*Toronto Star*$]$$[_{\text{VP}}$ reports $[_{\text{NP}}$ this$]$ $[_{\text{PP}}$ in the softest possible way$]$</a>,$[_{\text{S}}$ stating …$]]]$

Syntax of Mark-Up: Common Constituents

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

$$\mathrm{S} \rightarrow \mathtt{N}\underline{\mathtt{P\ V}}\mathtt{P}$$

Syntax of Mark-Up: Common Constituents

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way]</a>,[$_S$ stating ...]]]

$$\mathrm{S} \rightarrow \mathtt{N}\underline{\mathtt{P\ V}}\mathtt{P} \rightarrow \mathtt{DT}\ \underline{\mathtt{NNP\ NNP\ VBZ\ NP\ PP}}\ \mathtt{S}$$

Syntax of Mark-Up: Constituent Productions

	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Mark-Up: Constituent Productions

	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Mark-Up: Constituent Productions

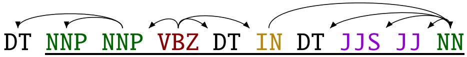
	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Mark-Up: Common Dependencies

..., but [$_{\text{S}}$ [$_{\text{NP}}$ the <a>*Toronto Star*]**[$_{\text{VP}}$ reports [$_{\text{NP}}$ this]**
[$_{\text{PP}}$ in the softest possible way]</a>**,[$_{\text{S}}$ stating ...]]]**

Syntax of Mark-Up: Common Dependencies

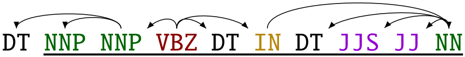
…, but $[_{\text{S}}$ $[_{\text{NP}}$ **the** <a>*Toronto Star*$]$$[_{\text{VP}}$ **reports** $[_{\text{NP}}$ **this**$]$
$[_{\text{PP}}$ **in the softest possible way**$]$</a>**,**$[_{\text{S}}$ **stating …**$]]]$

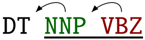


DT NNP NNP VBZ DT IN DT JJS JJ NN

Syntax of Mark-Up: Common Dependencies

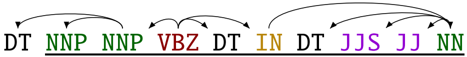
..., but [$_{S}$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way]</a>,[$_{S}$ stating ...]]]



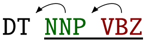


Syntax of Mark-Up: Common Dependencies

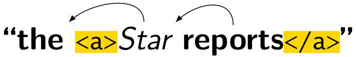
..., but [$_{S}$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way]</a>,[$_{S}$ stating ...]]]



DT NNP NNP VBZ DT IN DT JJS JJ NN



DT NNP VBZ

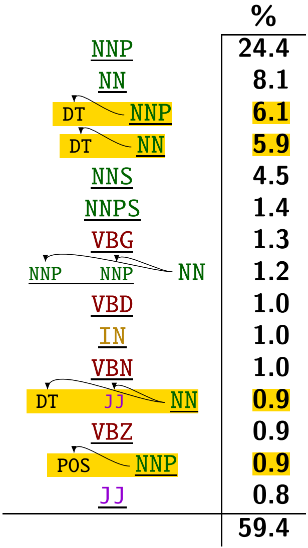


“the <a>*Star* reports</a>”

Syntax of Mark-Up: Head-Outward Spawns

	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Syntax of Mark-Up: Head-Outward Spawns

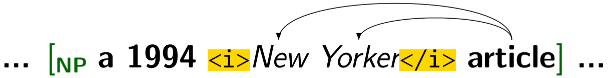


	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Syntax of Mark-Up: Head-Outward Spawns

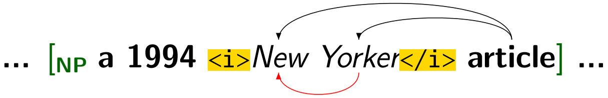
	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Syntax of Mark-Up: Exception



… [$_{\text{NP}}$ **a 1994** <i>*New Yorker*</i> **article**] **…**

Syntax of Mark-Up: Exception



… [$_{\text{NP}}$ 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 [$_{NP}$<a>[$_{NP}$ an analysis]</a>[$_{PP}$ 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>[$_{\text{PP}}$ 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 $[_{NP}$<a>$[_{NP}$ an analysis $]$</a>$[_{PP}$ 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 [$_{NP}$<a>[$_{NP}$ an analysis]</a>[$_{PP}$ 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)

Proposed Constraints: Constituents?

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

... in [$_{NP}$<a>[$_{NP}$ an analysis]</a>[$_{PP}$ 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

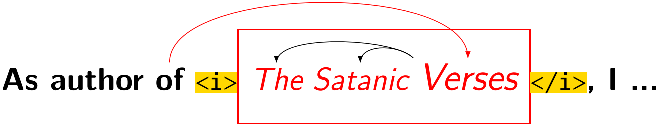
Proposed Constraints: Strict

Proposed Constraints: Strict

- **seal mark-up into attachments**

Proposed Constraints: Strict

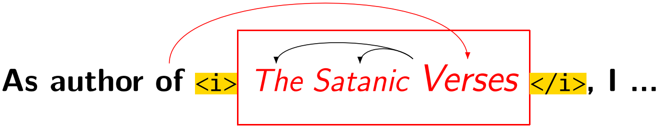
- **seal mark-up into attachments, e.g.,**



As author of `<i>` *The Satanic Verses* `</i>`**, I ...**

Proposed Constraints: Strict

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



As author of `<i>` *The Satanic Verses* `</i>`, I ...

— just 35.6% agreement with head-percolated trees

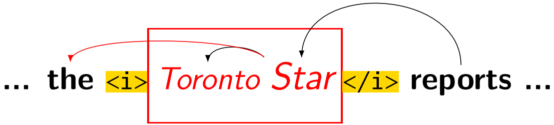
Proposed Constraints: Loose

Proposed Constraints: Loose

- **allow bracketed head word external dependents**

Proposed Constraints: Loose

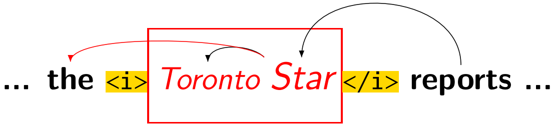
- **allow bracketed head word external dependents, e.g.,**



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

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

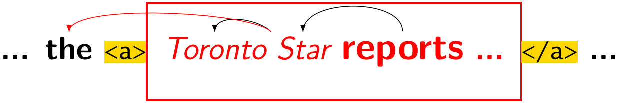
Proposed Constraints: Sprawl

Proposed Constraints: Sprawl

- **allow all bracketed words external dependents**

Proposed Constraints: Sprawl

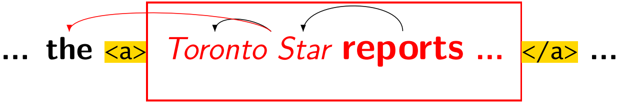
- **allow all bracketed words external dependents, e.g.,**



… 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

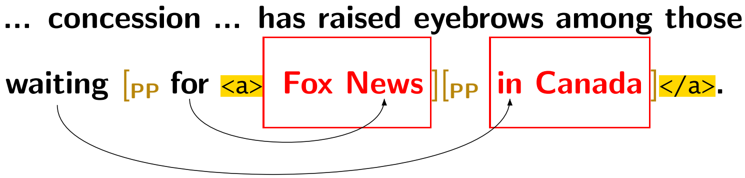
Proposed Constraints: Tear

Proposed Constraints: Tear

- **fracture by same-side external heads**

Proposed Constraints: Tear

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



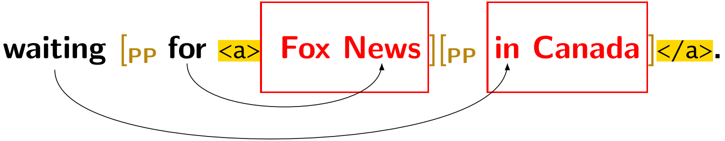
… concession … has raised eyebrows among those

waiting [$_{PP}$ for <a> Fox News][$_{PP}$ in Canada]</a>.

Proposed Constraints: Tear

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

… concession … has raised eyebrows among those



waiting [$_{\text{PP}}$ for <a> Fox News][$_{\text{PP}}$ in Canada]</a>.

— 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 (%)

Incarnation	**WSJ10**	**WSJ$^\infty$**	
(Cohen and Smith, 2009)			**Brown100**
(Spitkovsky et al., 2010)			
(Headden et al., 2009)			
BLOG			

Experimental Results: Dependency Accuracy (%)

Incarnation	WSJ10	WSJ$^{\infty}$	Brown100
(Cohen and Smith, 2009)	62.0	42.2	
(Spitkovsky et al., 2010)	57.1	45.0	43.6
(Headden et al., 2009)	68.8		
BLOG			

Experimental Results: Dependency Accuracy (%)

Incarnation	**WSJ10**	**WSJ$^{\infty}$**	**Brown100**
(Cohen and Smith, 2009)	**62.0**	**42.2**	
(Spitkovsky et al., 2010)	**57.1**	**45.0**	**43.6**
(Headden et al., 2009)	**68.8**		
BLOG	**69.3**	**50.4**	**53.3**

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

Experimental Results: Dependency Accuracy (%)

Incarnation	**WSJ10**	**WSJ$^{\infty}$**	**Brown100**
(Cohen and Smith, 2009)	**62.0**	**42.2**	
(Spitkovsky et al., 2010)	**57.1**	**45.0**	**43.6**
(Headden et al., 2009)	**68.8**		
BLOG	**69.3**	**50.4**	**53.3**
	$^{+}$**0.5**	$^{+}$**5.4**	$^{+}$**9.7**

- **state-of-the-art results**
- **linguistic constraints help with the task!**

Experimental Results: Dependency Accuracy (%)

	WSJ10	WSJ$^{\infty}$	Brown100
BLOG	**69.3**	**50.4**	**53.3**
NEWS			
WEB			

Experimental Results: Dependency Accuracy (%)

	WSJ10	WSJ$^\infty$	Brown100
BLOG	69.3	50.4	53.3
NEWS	67.3	50.1	51.6
WEB			

- **no need to manually clean data!**

Experimental Results: Dependency Accuracy (%)

	WSJ10	WSJ$^\infty$	Brown100
BLOG	69.3	50.4	53.3
NEWS	67.3	50.1	51.6
WEB	64.1	46.3	46.9

- **no need to manually clean data!**
- **nevertheless, less is more...**

Experimental Results: Dependency Accuracy (%)

	WSJ10	WSJ$^{\infty}$	Brown100
BLOG	69.3	50.4	53.3
NEWS	67.3	50.1	51.6
WEB	64.1	46.3	46.9

- no need to manually clean data!
- nevertheless, less is more...
- loose constraint consistently delivers best results

Experimental Results: Dependency Accuracy (%)

	WSJ10	WSJ$^{\infty}$	Brown100
BLOG	69.3	50.4	53.3
NEWS	67.3	50.1	51.6
WEB	64.1	46.3	46.9

- 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 (%)

	WSJ10	WSJ^{∞}	Brown100
BLOG	69.3	50.4	53.3
NEWS	67.3	50.1	51.6
WEB	64.1	46.3	46.9

- 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:**

	POS Sequence	**WEB** *Count*
	Sample web sentence, chosen uniformly at random.	
1	DT NNS VBN	**82,858,487**
	All rights reserved.	
2	NNP NNP NNP	**65,889,181**
	Yuasa et al.	
3	NN IN TO VB RB	**31,007,783**
	Sign in to YouTube now!	
4	NN IN IN PRP$ JJ NN	**31,007,471**
	Sign in with your Google Account!	

Experimental Results: Why Didn't the Web Help?

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

	POS Sequence	**WEB** *Count*
	Sample web sentence, chosen uniformly at random.	
1	DT NNS VBN	**82,858,487**
	All rights reserved.	
2	NNP NNP NNP	**65,889,181**
	Yuasa et al.	
3	NN IN TO VB RB	**31,007,783**
	Sign in to YouTube now!	
4	NN IN IN PRP$ JJ NN	**31,007,471**
	Sign in with your Google Account!	

Experimental Results: Why Didn't the Web Help?

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

	POS Sequence	**WEB** *Count*
	Sample web sentence, chosen uniformly at random.	
1	DT NNS VBN	**82,858,487**
	All rights reserved.	
2	NNP NNP NNP	**65,889,181**
	Yuasa et al.	
3	NN IN TO VB RB	**31,007,783**
	Sign in to YouTube now!	
4	NN IN IN PRP$ JJ NN	**31,007,471**
	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:**

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

Potential

- **another motivating example:**

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

  — internal structure of a compound

  (Vadas and Curran, 2007)

Potential

- **another motivating example:**

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

  — **internal structure of a compound**

  (Vadas and Curran, 2007)

  — **lower-level tokenization signal**

Potential

- **another motivating example:**

[$_{NP}$ [$_{NP}$ **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/

(NP (ADJP (NP (JJ Libyan) (NN ruler))
(JJ Mu))
(“ ‘) (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?**

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

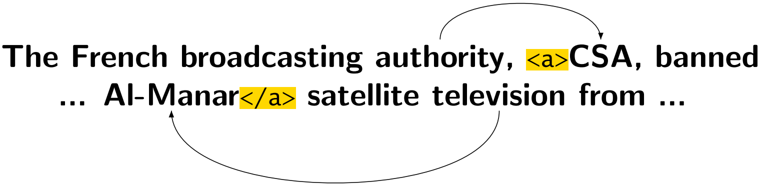
- **remaining 1.1% mostly due to parser errors...**

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, <a>CSA, banned
... Al-Manar</a> satellite television from ...