Any Domain Parsing
Automatic Domain Adaptation for Natural Language Parsing

David McClosky

Department of Computer Science
Brown University

September 18th, 2009

Committee: Eugene Charniak, Mark Johnson, and Dan Klein
Robots will need to understand language
Robots will need to understand language

Keeping up to date with Twitter
Rebel Alliance
Associated Press

Rebels make Death Star go Nova!
by Hon. Princess Leia (RAP writer), Hans Solo (RAP contributor)

DEATH STAR -- At 3:22pm Galactic Central Time, Rebel fighters launched an assault which ultimately lead to the destruction of
Studying the latest medical journals

Journal of Prosthetics and Cybernetics
Volume VI, Issue IV

Robotic hand grafted after "lightsaber accident"
Casual reading

A TALE OF TWO FORCES

ZEN AND THE ART OF LIGHTSABER MAINTENANCE
Imperial Senate left waffles on Death Star.
Semantics depend on syntax
Semantics depend on syntax
Semantics depend on syntax

\[ S \]

\[ NP \]

- **Adj**: Imperial
- **Noun**: Senate
- **Noun**: left

\[ VP \]

- **Verb**: waffles

\[ PP \]

- **Prep**: on
- **NP**
  - **Noun**: Death
  - **Noun**: Star

---

Drawing of a planet with labeled parts:

- Front View
- Surface City Blocks
- Superlaser
  - Focus Lens
- Equatorial Trench
- Exposed Superstructure
- Command Sector (South)
- Emperor's Tower
- Command Sector (North)
- Reactor Core (internal)
- Ion Drives (uncompleted)
Semantics depend on syntax

```
S
  NP
    Adj    Noun
    Imperial Senate
  VP
    Verb
    left
    NP
    Noun
    waffles
  PP
    Prep
    on
    NP
    Noun
    Death
    Noun
    Star
```
Applications of Parsing

- Parsing is often part of larger NLP pipelines.
- Some examples:
  - Machine translation [Charniak et al., 2003]
  - Bioinformatics [Miyao et al., 2008]
  - Forensics (author identification) [Luyckx and Daelemans, 2008]
  - Discourse analysis [Barzilay and Lapata, 2008]
  - Summarization [Turner and Charniak, 2005]
  - Language modeling [Roark, 2001], [Charniak, 2001]
  - Speech repairs [Johnson and Charniak, 2004]
  - Coreference [Luo and Zitouni, 2005], [Charniak and Elsner, 2009]
  - etc.
Many current approaches to parsing are data-driven.
Data-driven Parsing

- Many current approaches to parsing are data-driven.
- Data consists of human-annotated corpora with labeled examples of correct parse structures ("gold trees").
Data-driven Parsing

- Parsers are trained on these corpora to produce models.

![Graph showing training process]
The parsing model is used to parse unlabeled text.
Data-driven Parsing

- Many model parameters may hurt portability/generality

train

gold trees
1 million words, 40k trees

parse

~2 million parameters

parsed text

raw text

Sequential development of structural and functional ... The TCR alpha beta or -gamma delta chains bind the ... Recently, several groups have described marked ... The sequence in which these alterations develop is ...
What's in a domain?
Masters of their domain?

true parse (gold)
Masters of their domain?

true parse (gold)

from newspaper article parser
Masters of their domain?

true parse (gold)

from newspaper article parser
Masters of their domain?

true parse (gold)

from newspaper article parser
Evaluating parse trees

true parse (gold)

from newspaper article parser
Evaluating parse trees

true parse (gold)

to
activate

from newspaper article parser
Evaluating parse trees

true parse (gold)

from newspaper article parser
Evaluating parse trees

**true parse (gold)**

\[
\text{recall} = \frac{2}{6}
\]

\[
f\text{-score} = \frac{2(\text{precision})(\text{recall})}{\text{precision} + \text{recall}} = \frac{4}{11}
\]

**from newspaper article parser**
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>supervised parsing</td>
</tr>
</tbody>
</table>
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Train Diagram]</td>
<td>![Test Diagram]</td>
<td>supervised parsing</td>
</tr>
<tr>
<td>![Train Diagram]</td>
<td>![Test Diagram]</td>
<td>semi-supervised parsing</td>
</tr>
</tbody>
</table>
# Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Train Scenario" /></td>
<td><img src="image2.png" alt="Test Scenario" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image3.png" alt="Train Scenario" /></td>
<td><img src="image4.png" alt="Test Scenario" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image5.png" alt="Train Scenario" /></td>
<td><img src="image6.png" alt="Test Scenario" /></td>
<td>parser portability</td>
</tr>
</tbody>
</table>
### Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>supervised parsing</td>
</tr>
<tr>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>parser portability</td>
</tr>
<tr>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>parser adaptation</td>
</tr>
</tbody>
</table>
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Train Tree" /></td>
<td><img src="image" alt="Test Tree" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image" alt="Train Tree" /></td>
<td><img src="image" alt="Test Tree" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image" alt="Train Tree" /></td>
<td><img src="image" alt="Test Tree" /></td>
<td>parser portability</td>
</tr>
<tr>
<td><img src="image" alt="Train Tree" /></td>
<td><img src="image" alt="Test Tree" /></td>
<td>parser adaptation</td>
</tr>
<tr>
<td><img src="image" alt="Train Tree" /></td>
<td><img src="image" alt="Test Tree" /></td>
<td>multiple source parser adaptation</td>
</tr>
<tr>
<td>Train</td>
<td>Test</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image3.png" alt="Diagram" /></td>
<td><img src="image4.png" alt="Diagram" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image5.png" alt="Diagram" /></td>
<td><img src="image6.png" alt="Diagram" /></td>
<td>parser portability</td>
</tr>
<tr>
<td><img src="image7.png" alt="Diagram" /></td>
<td><img src="image8.png" alt="Diagram" /></td>
<td>parser adaptation</td>
</tr>
<tr>
<td><img src="image9.png" alt="Diagram" /></td>
<td><img src="image10.png" alt="Diagram" /></td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Train Diagram 1]</td>
<td>![Test Diagram 1]</td>
<td>supervised parsing</td>
</tr>
<tr>
<td>![Train Diagram 2]</td>
<td>![Test Diagram 2]</td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td>![Train Diagram 3]</td>
<td>![Test Diagram 3]</td>
<td>parser portability</td>
</tr>
<tr>
<td>![Train Diagram 4]</td>
<td>![Test Diagram 4]</td>
<td>parser adaptation</td>
</tr>
<tr>
<td>![Train Diagram 5]</td>
<td>![Test Diagram 5]</td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Thesis statement

Self-training is an effective semi-supervised learning technique for parsing, capable of improving both in-domain and cross-domain parsing scenarios.
<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Train Diagram" /></td>
<td><img src="image2" alt="Test Diagram" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image3" alt="Train Diagram" /></td>
<td><img src="image4" alt="Test Diagram" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image5" alt="Train Diagram" /></td>
<td><img src="image6" alt="Test Diagram" /></td>
<td>parser portability</td>
</tr>
<tr>
<td><img src="image7" alt="Train Diagram" /></td>
<td><img src="image8" alt="Test Diagram" /></td>
<td>parser adaptation</td>
</tr>
<tr>
<td><img src="image9" alt="Train Diagram" /></td>
<td><img src="image10" alt="Test Diagram" /></td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Incorporating unlabeled data

- How can we leverage unlabeled data in our models?

```
Sequential development of structural and functional ...
The TCR alpha beta or -gamma delta chains bind the ...
Recently, several groups have described marked ...
The sequence in which these alterations develop is ...
```

`raw text`
Incorporating unlabeled data: Self-training

1. Train a model from the labeled data.
Incorporating unlabeled data: Self-training

2. Parse the unlabeled text.

Sequential development of structural and functional ... The TCR alpha beta or gamma delta chains bind the ... Recently, several groups have described marked ... The sequence in which these alterations develop is ...
Incorporating unlabeled data: Self-training

3. Combine gold trees with automatically parsed trees.
4. Train a new model from the combination.

Incorporating unlabeled data: Self-training
A Brief History of Self-training

- [Charniak, 1997]
- [Steedman et al., 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]
A Brief History of Self-training

- [Charniak, 1997]
- [Steedman et al., 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]

→ no improvements over state-of-the-art from self-training
A Brief History of Self-training

- [Charniak, 1997]
- [Steedman et al., 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]
  → no improvements over state-of-the-art from self-training
- [McClosky, Charniak, and Johnson, NAACL 2006]
A Brief History of Self-training

- [Charniak, 1997]
- [Steedman et al., 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]
  → no improvements over state-of-the-art from self-training
- [McClosky, Charniak, and Johnson, NAACL 2006]
  → reranking parser
The Parsing Model
[Charniak and Johnson, 2005]

- Lexicalized PCFG parser gives most probable parse

And now for something completely different...

sentence

parsing model

parse

“first stage parser”
or
“parser”

tree
The Parsing Model

[Charniak and Johnson, 2005]

- Use $n$ most probable parses instead just top parse

And now for something completely different...

sentence

$parsing \ model$

$n$-best

$parse_1$

$n$ trees
The Parsing Model
[Charniak and Johnson, 2005]

- Discriminative reranker picks “best” parse from list
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>supervised parsing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>parser portability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>parser adaptation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Self-training for parsing

Gold trees \rightarrow \text{self-train} \rightarrow \text{self-trained parsing model} \rightarrow \text{parse} \rightarrow \text{test trees}

Raw text

Sequential development of structural and functional... The TCR alpha beta or gamma delta chains bind the...
Recently, several groups have described marked...
The sequence in which these alterations develop is...
Self-training for parsing

Training:
- **WSJ**
  - Gold trees
  - Wall Street Journal
    - 40,000 trees
    - Newspaper articles
  - NANC
    - North American News Text Corpus
    - 2 million sentences
    - Newspaper articles
- **Self-train**
- **Self-trained parsing model**

Evaluation:
- **Wall Street Journal**
- **Parse**
- **Test sentences**
- **Test trees**
Self-training for parsing

**Training**
- gold trees
- self-train
  - use parser or reranking parser?
- self-trained parsing model

**Evaluation**
- raw text
- test sentences
- parse
  - (reranking parser)
- test trees
Self-training for parsing is effective
[McClosky, Charniak, and Johnson, NAACL 2006]

<table>
<thead>
<tr>
<th>Model</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (WSJ)</td>
<td>91.3</td>
</tr>
<tr>
<td>Self-trained (WSJ + NANC)</td>
<td>92.1</td>
</tr>
</tbody>
</table>

*f*-scores on WSJ evaluation section
## Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Train Diagram" /></td>
<td><img src="image2" alt="Test Diagram" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image3" alt="Train Diagram" /></td>
<td><img src="image4" alt="Test Diagram" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image5" alt="Train Diagram" /></td>
<td><img src="image6" alt="Test Diagram" /></td>
<td>parser portability</td>
</tr>
<tr>
<td><img src="image7" alt="Train Diagram" /></td>
<td><img src="image8" alt="Test Diagram" /></td>
<td>parser adaptation</td>
</tr>
<tr>
<td><img src="image9" alt="Train Diagram" /></td>
<td><img src="image10" alt="Test Diagram" /></td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Parser portability experiments

---

Training:
- **WSJ** gold trees
- **self-train**
- **self-trained parsing model**

Evaluation:
- **roll** test sentences
- **parse**
- **test trees**

---

Sequential development of structural and functional ---
The TCR alpha beta ---
Recently, several ---
Described marked ---
The sequence in which ---
Rolls-Royce Motors ---
The luxury automaker ---
Howard Mosher, ---
Bell, based in Los ---

---

33
Parser portability experiments

**Training**

- **gold trees**
- **self-train**
- **self-trained parsing model**

**Evaluation**

- **Brown Corpus**
  - Mixture of several non-news domains
- **parse**
- **test sentences**
- **test trees**
Self-trained WSJ model portability

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>WSJ</td>
<td>91.3</td>
</tr>
<tr>
<td>WSJ</td>
<td>BROWN</td>
<td>85.2</td>
</tr>
</tbody>
</table>

*f*-score on WSJ and BROWN evaluation sections
### Self-trained WSJ model portability

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>WSJ</td>
<td>91.3</td>
</tr>
<tr>
<td>WSJ</td>
<td>BROWN</td>
<td>85.2</td>
</tr>
<tr>
<td>WSJ + NANC</td>
<td>BROWN</td>
<td>87.8</td>
</tr>
</tbody>
</table>

*f-score on WSJ and BROWN evaluation sections*
## Self-trained WSJ model portability

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>$f$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>WSJ</td>
<td>91.3</td>
</tr>
<tr>
<td>WSJ</td>
<td>BROWN</td>
<td>85.2</td>
</tr>
<tr>
<td>WSJ + NANC</td>
<td>BROWN</td>
<td>87.8</td>
</tr>
<tr>
<td>BROWN</td>
<td>BROWN</td>
<td>88.4</td>
</tr>
</tbody>
</table>

$f$-score on WSJ and BROWN evaluation sections
Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Diagram 1]</td>
<td>![Diagram 2]</td>
<td>supervised parsing</td>
</tr>
<tr>
<td>![Diagram 3]</td>
<td>![Diagram 4]</td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td>![Diagram 5]</td>
<td>![Diagram 6]</td>
<td>parser portability</td>
</tr>
<tr>
<td>![Diagram 7]</td>
<td>![Diagram 8]</td>
<td>parser adaptation</td>
</tr>
<tr>
<td>![Diagram 9]</td>
<td>![Diagram 10]</td>
<td>multiple source parser adaptation</td>
</tr>
</tbody>
</table>
Parser adaptation experiments

More distant domains...

Training

- **WSJ**
- **gold trees**
- **NANC**
- **raw text**

Training process:

1. Self-train
2. Create self-trained parsing model

Evaluation

- **Brown Corpus**: Mixture of several non-news domains
- **test sentences**

Evaluation process:

1. Parse
2. Test trees

Sequential development of structural and functional properties

Recently, several examples of described marked genes in which our procedures develop
Parser adaptation experiments

More distant domains...

Training:
- WSJ gold trees
- NANC raw text

Self-train

Self-trained parsing model

Evaluation:
- GENIA Corpus: Biomedical journal article abstracts
- GENIA test sentences

Parse

Test trees
Parser adaptation experiments
More distant domains...

Training:
- gold trees
- WSJ
- PubMed

Self-train

Self-trained parsing model

Evaluation:
- GENIA Corpus
  Biomedical journal article abstracts
- test sentences

Parse

raw text
- PubMed
  270,000 sentences
  Biomedical journal article abstracts

NANC
- Sequential development of structural ar
  r-gamma delta cri
  ups have described
  The sequence in which these alterations...
Parser adaptation experiments
More distant domains...

- **Training**:
  - WSJ gold trees
  - PubMed raw text

- **Self-train**
- **Self-trained parsing model**

- **Evaluation**:
  - GENIA Corpus
    - Biomedical journal article abstracts
  - GENIA test sentences

- **Output**:
  - parse
  - test trees

- **BioBooks**
  - 70,000 sentences
  - Biology textbooks scraped from web
Varying unlabeled data for self-training

[McClosky and Charniak, ACL 2008]

$f$-score on GENIA development section
## Parsing training scenarios

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Tree Diagram 1" /></td>
<td><img src="image2.png" alt="Tree Diagram 2" /></td>
<td>supervised parsing</td>
</tr>
<tr>
<td><img src="image3.png" alt="Tree Diagram 3" /></td>
<td><img src="image4.png" alt="Tree Diagram 4" /></td>
<td>semi-supervised parsing</td>
</tr>
<tr>
<td><img src="image5.png" alt="Tree Diagram 5" /></td>
<td><img src="image6.png" alt="Tree Diagram 6" /></td>
<td>parser portability</td>
</tr>
<tr>
<td><img src="image7.png" alt="Tree Diagram 7" /></td>
<td><img src="image8.png" alt="Tree Diagram 8" /></td>
<td>parser adaptation</td>
</tr>
</tbody>
</table>

![Multiple Source Tree](image9.png) | ? | multiple source parser adaptation |
Automatic Domain Adaptation

- What if we don’t know the target domain?
Automatic Domain Adaptation

- What if we don’t know the target domain?
  - Parsing the web or any other large heterogeneous corpus
Automatic Domain Adaptation

- What if we don’t know the target domain?
  - Parsing the web or any other large heterogeneous corpus
- Consider a new parsing task:
  - labeled and unlabeled corpora (source domains)
Automatic Domain Adaptation

- What if we don’t know the target domain?
  - Parsing the web or any other large heterogeneous corpus
- Consider a new parsing task:
  - labeled and unlabeled corpora (source domains)
  - documents to parse (target text)
Any Domain Parsing

source domain models

target text
Crossdomain Accuracy Prediction
Crossdomain Accuracy Prediction

- [WSJ] 89%
- [Brown] 83%
- [WSJ] 90%
- [WSJ] 84%
Crossdomain Accuracy Prediction

- WSJ $\rightarrow$ WSJ: 89%
- Brown $\rightarrow$ WSJ: 83%
- [Colors] $\rightarrow$ WSJ: 90%
- [Colors] $\rightarrow$ WSJ: 84%
Crossdomain Accuracy Prediction

predict(?, WSJ) = f-score

Similar to [Ravi et al., 2008]
Prediction by regression

\[ \text{predict}(), \hat{y} \] = \text{f-score} (predicted)
Regression features

\[ \text{predict}(\text{domainagnostic}), \text{?} ) = f\text{-score} \]

(predicted)

Domain Divergence Measures
Regression features

\[ \text{predict}(), \quad \text{(predicted)} = f\text{-score} \]

Domain Divergence Measures
Regression features

\[
predict(\text{\begin{array}{c}1\end{array}}, \text{\begin{array}{c}2\end{array}}) = f\text{-score}
\]
(predicted)

Domain Divergence Measures
Regression features

\[ \text{predict}(\text{, } ?) = \]

Domain Divergence Measures

Divide mixture weight by divergence:
Cosine Similarity

, the . of and

☐ ☐ ☐ ☐ ☐
Cosine Similarity

<table>
<thead>
<tr>
<th>,</th>
<th>the</th>
<th>.</th>
<th>of</th>
<th>and</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9%</td>
<td>5.1%</td>
<td>3.8%</td>
<td>2.4%</td>
<td>1.8%</td>
</tr>
<tr>
<td>3.6%</td>
<td>4.6%</td>
<td>3.6%</td>
<td>4.2%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>
Cosine Similarity

, the . of and

4.9%  5.1%  3.8%  2.4%  1.8%

3.6%  4.6%  3.6%  4.2%  2.6%

\[
\text{cosine similarity} = \frac{\langle \text{WSJ}, \text{GENIA} \rangle}{\| \text{WSJ} \| \cdot \| \text{GENIA} \|} \approx 0.956
\]
Unknown words

WSJ

FAKE NEWS

GOOD NEWS

BAD NEWS
Unknown words

WSJ

FAKE NEWS
GOOD NEWS
BAD NEWS

= vocabulary
Unknown words

WSJ

GENIA

= vocabulary
Unknown words

= vocabulary
Regression features

\[ \text{predict}(\text{domain}, \text{?}) = f\text{-score} \]

Domain Divergence Measures
Regression features

\[
predict(\text{Source domain features}, \text{?}) = f\text{-score (predicted)}
\]
Regression features

\[ \text{predict}(\text{source domain mixture}, \text{uniform}) = f\text{-score (predicted)} \]
Regression features

\[
predict(?, \text{Source domain features}) = f\text{-score (predicted)}
\]

Entropy:
\[
H(X) = - \sum_{i=1}^{n} P(x_i) \log P(x_i)
\]
Model and estimation

\[
predict(\cdot, \cdot) = \vec{\lambda}_c \cosinesim(\cdot, \cdot) + \vec{\lambda}_{\vec{w}} \unkwords(\cdot, \cdot) + \vec{\lambda}_u \entropy(\cdot) + b
\]
Model and estimation

\[
\text{predict}(\vec{\lambda}_c, \text{cosinesim}(\vec{\lambda}_c, \text{unkwords}(\vec{\lambda}_w, \text{entropy}(\vec{\lambda}_u))) + b)
\]
Anatomy of a data point

**input**

- **source**
- **target**
- **cosinesim**: 0.94
- **unkwords**: 27%
- **entropy**: 3.83 bits

**output**

- **f-score**: 78%

* numbers on this slide are cooked
Training data

* numbers on this slide are cooked
Round-robin evaluation
Round-robin evaluation
Evaluation for GENIA

**train**

**sources**

- WSJ
- Brown
- ETT
- NANC
- Gutenberg

**targets**

- WSJ
- Brown
- ETT
- SWITCHBOARD
- BNC
Evaluation for GENIA

**train**

**sources**

- WSJ
- Brown
- NANC
- Gutenberg

**targets**

- WSJ
- Brown
- ETT
- BNC

**test**

**sources**

- WSJ
- Brown
- NANC
- Gutenberg

**target**

- GENIA
Baselines

- Standard baselines
Baselines

- Standard baselines
  - Fixed set: WSJ
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)

- Oracle baselines
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)

- Oracle baselines
  - Best single corpus
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)

- Oracle baselines
  - Best single corpus
  - Best seen
Moral of the story

- What’s the best way to parse new text?
  - Self-training on similar text improves performance
  - Any Domain Parsing provides additional benefits by selecting relevant corpora

- Self-training helps in many different parsing scenarios
  - Allows us to use unlabeled data to improve performance
  - State-of-the-art performance on WSJ, BROWN, and GENIA

- Relevant publications:
  - [McClosky, Charniak, and Johnson, NAACL 2006]
  - [McClosky, Charniak, and Johnson, ACL 2006]
  - [McClosky and Charniak, ACL 2008]
  - [McClosky, Charniak, and Johnson, COLING 2008]
May The Force Be With You

Questions?

Thanks to my committee, BLLIP, friends, and family for their feedback and support!

Dedicated to my grandparents

Brought to you by NSF grants LIS9720368 and IIS0095940 and DARPA GALE contract HR0011-06-2-0001
Extra slides
What is...the matrix?

<table>
<thead>
<tr>
<th>Train</th>
<th>Literature</th>
<th>BioMed</th>
<th>Phone</th>
<th>ETT</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>86.7</td>
<td>73.5</td>
<td>77.6</td>
<td>80.8</td>
<td>79.9</td>
</tr>
<tr>
<td>BioMed</td>
<td>65.7</td>
<td>84.6</td>
<td>50.5</td>
<td>67.1</td>
<td>64.6</td>
</tr>
<tr>
<td>Phone</td>
<td>75.8</td>
<td>63.6</td>
<td>88.2</td>
<td>76.2</td>
<td>69.8</td>
</tr>
<tr>
<td>ETT</td>
<td>76.2</td>
<td>65.7</td>
<td>74.5</td>
<td>82.4</td>
<td>72.6</td>
</tr>
<tr>
<td>News</td>
<td>84.1</td>
<td>76.2</td>
<td>76.7</td>
<td>82.2</td>
<td>89.7</td>
</tr>
</tbody>
</table>

(f-scores on all sentences in test sets)
The Four Hypotheses
[McClosky, Charniak, and Johnson, COLING 2008]

Four hypotheses:

1. Self-training works after a phase transition.
2. Self-trained parser makes fewer search errors.
4. Self-training teaches the parser about bilexical dependencies.
In-domain evaluation

**train**

**sources**

- WSJ
- Brown
- ETT
- GENIA
- NANC
- Gutenberg

**targets**

- WSJ
- Brown
- ETT
- GENIA
- PubMed

**test**

**sources**

- WSJ
- Brown
- ETT
- GENIA
- NANC
- Gutenberg

**target**

- GENIA
In-domain evaluation results
Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length
Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length
  - # of conjunctions
Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length
  - # of conjunctions
  - # of bigrams in NANC not seen in WSJ
Why does self-training help?

[Hypothesis: Self-training teaches the parser about bilexical dependencies.]

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]
The Generative Story

```
  .
  .
VP
  |
Verb
  |
  were

  |
Adjp
  |
Adj
  |
occupying

  |
NP
```
The Generative Story

```
        ...
        ...
VP

Verb  ADJP

Adj  label  NP

were  occupying

```
The Generative Story

VP
  Verb
  were
  occupying
  ADJP
  Adj
  label
  NP

history

?
The Generative Story

```
Verb
  
were

Adj

occupying

ADJP

history

NP

label

head tag

Noun
```
The Generative Story

```
history
```

```
<table>
<thead>
<tr>
<th>VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
</tr>
<tr>
<td>were</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>ADJP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj</td>
</tr>
<tr>
<td>occupying</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>head</td>
</tr>
<tr>
<td>tag</td>
</tr>
<tr>
<td>building</td>
</tr>
</tbody>
</table>

```

```
| head word |
```


The Generative Story

Diagram of a sentence:

- **Verb**: were
- **Adj**: occupying
- **ADJP**: occupying
- **NP**: building

**Bilexical Dependency**:

- **Label**: history
- **Head Tag**: head word
The Generative Story

```
VP
  Verb
    were
  ADJP
    Adj
      occupying
    NP
      Det
        a
      Noun
        university
      Noun
        building
```

- **history**: label
- **head tag**: head word
- **expansion**
The Generative Story

![Grammar Tree Diagram]

- **Verb**: were
- **ADJP**
  - **Adj**: occupying
  - **NP**: a university building
What does self-training teach the parser?
[McClosky, Charniak, and Johnson, COLING 2008]

\[
P(\text{constit} \mid \text{label}, \text{history}) = P(\text{tag} \mid \text{label}, \text{history}) \\
\times P(\text{head} \mid \text{tag}, \text{label}, \text{history}) \\
\times P(\text{exp} \mid \text{head}, \text{tag}, \text{label}, \text{history})
\]
What does self-training teach the parser?

[McClosky, Charniak, and Johnson, COLING 2008]

\[ P(\text{constit} \mid \text{label}, \text{history}) = P(\text{tag} \mid \text{label}, \text{history}) \times P(\text{head} \mid \text{tag, label, history}) \times P(\text{exp} \mid \text{head, tag, label, history}) \]
What does self-training teach the parser?

[McClosky, Charniak, and Johnson, COLING 2008]

\[ P(\text{constit} \mid \text{label}, \text{history}) = P(\text{tag} \mid \text{label}, \text{history}) \times P(\text{head} \mid \text{tag}, \text{label}, \text{history}) \times P(\text{exp} \mid \text{head}, \text{tag}, \text{label}, \text{history}) \]
What we learn from unlabeled data
What we learn from unlabeled data
What we learn from unlabeled data
What we learn from unlabeled data
Outline

Introduction
  Parsing basics and applications
  Domain Dependence

Self-training for Parsing
  Self-training
  Semi-supervised parsing
  Parser portability
  Parser adaptation

Any Domain Parsing: Automatic Domain Adaptation
  Any Domain Parsing
  Estimation and training data
  Evaluation
  Results

Conclusion

Extra slides
  In-domain evaluation
  Analysis