Automatic Domain Adaptation for Parsing

David McClosky\textsuperscript{a,b}  Eugene Charniak\textsuperscript{b}  Mark Johnson\textsuperscript{c,b}

\textsuperscript{a} Natural Language Processing Group
Stanford University

\textsuperscript{b} Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University

\textsuperscript{c} Department of Computing
Macquarie University

(work performed while all authors were at Brown)

NAACL-HLT 2010 — June 2nd, 2010
Understanding 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
What’s in a domain?
## Crossdomain parsing performance

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>WSJ</td>
</tr>
</tbody>
</table>

89.7

(f-scores on all sentences in test sets, Charniak parser)
Crossdomain parsing performance

<table>
<thead>
<tr>
<th>Train</th>
<th>BROWN</th>
<th>GENIA</th>
<th>SWBD</th>
<th>ETT</th>
<th>WSJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>BROWN</td>
<td>86.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENIA</td>
<td></td>
<td>84.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWBD</td>
<td></td>
<td></td>
<td>88.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETT</td>
<td></td>
<td></td>
<td></td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>WSJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.7</td>
</tr>
</tbody>
</table>

(f-scores on all sentences in test sets, Charniak parser)
Crossdomain parsing performance...not great

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BROWN</td>
</tr>
<tr>
<td>BROWN</td>
<td>86.7</td>
</tr>
<tr>
<td>GENIA</td>
<td>65.7</td>
</tr>
<tr>
<td>SWBD</td>
<td>75.8</td>
</tr>
<tr>
<td>ETT</td>
<td>76.2</td>
</tr>
<tr>
<td>WSJ</td>
<td>84.1</td>
</tr>
</tbody>
</table>

(f-scores on all sentences in test sets, Charniak parser)
Color key: < 70, 70–80, > 80
What if we don’t know the target domain?
Automatic Domain Adaptation for Parsing

- What if we don’t know the target domain?
  - Parsing the web or any other large heterogeneous corpus
What if we don’t know the target domain?
  ▶ Parsing the web or any other large heterogeneous corpus
  ▶ A new hope parsing task:
Automatic Domain Adaptation for Parsing

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

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

A new hope parsing task:
  labeled and unlabeled corpora (source domains)
  corpora to parse (target text)

Combine source domains to best parse each target text
What if we don’t know the target domain?
  - Parsing the web or any other large heterogeneous corpus

A new hope parsing task:
  - labeled and unlabeled corpora (source domains)
  - corpora to parse (target text)

Combine source domains to best parse each target text

Evaluation: parse unknown and foreign domains
Related work

- **Subdomain Sensitive Parsing**  
  [Plank and Sima’an, LREC 2008]
  - Extract subdomains from WSJ using domain-specific LMs
  - Use above to train domain-specific parsing models

- **Multitask learning**  
  [Daumé III, 2007], [Finkel and Manning, 2009]
  - Each domain is a separate (related) task
  - Share non-domain specific information across domains

- **Predicting parsing performance**  
  [Ravi, Knight, and Soricut, EMNLP 2008]
  - Use regression to predict $f$-score of a parse
  - Predicted accuracies can be used to rank models
Crossdomain accuracy prediction
Crossdomain accuracy prediction

- WSJ → WSJ: 89%
- Brown → WSJ: 83%
- [Colors] → WSJ: 90%
- [Colors] → WSJ: 84%
Crossdomain accuracy prediction

\[ \text{predict}(\square, \hexagon) = f\text{-score} \]
Prediction by regression

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

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

Domain Divergence Measures
Regression features

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

Domain Divergence Measures
Regression features

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

Domain Divergence Measures
Regression features

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

Domain Divergence Measures

Divide mixture weight by divergence:
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%

WSJ
GENIA
Cosine Similarity

, the . of and

\[
\begin{array}{ccccc}
4.9\% & 5.1\% & 3.8\% & 2.4\% & 1.8\% \\
3.6\% & 4.6\% & 3.6\% & 4.2\% & 2.6\%
\end{array}
\]

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

= vocabulary
Regression features

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

$$\text{predict}(\text{Source domain features}) = f\text{-score (predicted)}$$
Regression features

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

Source domain features
Regression features

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

Entropy:

\[
H(X) = - \sum_{i=1}^{n} P(x_i) \log P(x_i)
\]
Features considered

- **Domain divergence measures**
  - $n$-gram language model (PPL, PPL1, probability)
  - Cosine similarity for frequent words
    - ($k \in \{5, 50, 500, 5000\}$)
  - Cosine similarity for punctuation
  - Average length differences (absolute, directed)
  - % unknown words (source $\rightarrow$ target, target $\rightarrow$ source)

- **Source domain features**
  - Source domain probabilities
  - Source domain non-zero probability
  - # source domains
  - % self-trained corpora
  - Source domain entropy
Features considered

- Domain divergence measures
  - $n$-gram language model (PPL, PPL1, probability)
  - **Cosine similarity for frequent words**
    - ($k \in \{5, 50, 500, 5000\}$)
  - Cosine similarity for punctuation
  - Average length differences (absolute, directed)
  - % **unknown words** (source $\rightarrow$ target, target $\rightarrow$ source)

- Source domain features
  - Source domain probabilities
  - Source domain non-zero probability
  - # source domains
  - % self-trained corpora
  - **Source domain entropy**
Cosine similarity illustrated ($k = 5000$)

<table>
<thead>
<tr>
<th>Source domain</th>
<th>BNC</th>
<th>GENIA</th>
<th>BROWN</th>
<th>SWBD</th>
<th>ETT</th>
<th>WSJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENIA</td>
<td>0.894</td>
<td>0.998</td>
<td>0.860</td>
<td>0.676</td>
<td>0.887</td>
<td>0.881</td>
</tr>
<tr>
<td>PUBMED</td>
<td>0.911</td>
<td>0.977</td>
<td>0.875</td>
<td>0.697</td>
<td>0.895</td>
<td>0.897</td>
</tr>
<tr>
<td>BROWN</td>
<td>0.976</td>
<td>0.862</td>
<td>0.999</td>
<td>0.828</td>
<td>0.917</td>
<td>0.960</td>
</tr>
<tr>
<td>GUTENBERG</td>
<td>0.982</td>
<td>0.868</td>
<td>0.977</td>
<td>0.839</td>
<td>0.929</td>
<td>0.957</td>
</tr>
<tr>
<td>SWBD</td>
<td>0.779</td>
<td>0.663</td>
<td>0.825</td>
<td>0.992</td>
<td>0.695</td>
<td>0.789</td>
</tr>
<tr>
<td>ETT</td>
<td>0.971</td>
<td>0.896</td>
<td>0.937</td>
<td>0.766</td>
<td>0.992</td>
<td>0.959</td>
</tr>
<tr>
<td>WSJ</td>
<td>0.968</td>
<td>0.880</td>
<td>0.963</td>
<td>0.803</td>
<td>0.941</td>
<td>0.997</td>
</tr>
<tr>
<td>NANC</td>
<td>0.983</td>
<td>0.888</td>
<td>0.979</td>
<td>0.801</td>
<td>0.950</td>
<td>0.987</td>
</tr>
</tbody>
</table>
## Unknown words illustrated (target $\rightarrow$ source)

<table>
<thead>
<tr>
<th>Source domain</th>
<th>BNC</th>
<th>GENIA</th>
<th>BROWN</th>
<th>SWBD</th>
<th>ETT</th>
<th>WSJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENIA</td>
<td>33.3</td>
<td>10.8</td>
<td>40.5</td>
<td>45.8</td>
<td>43.1</td>
<td>38.9</td>
</tr>
<tr>
<td>PUBMED</td>
<td>32.5</td>
<td>21.5</td>
<td>36.5</td>
<td>45.4</td>
<td>42.0</td>
<td>35.5</td>
</tr>
<tr>
<td>BROWN</td>
<td>14.3</td>
<td>38.5</td>
<td>10.7</td>
<td>21.5</td>
<td>22.7</td>
<td>18.3</td>
</tr>
<tr>
<td>GUTENBERG</td>
<td>16.0</td>
<td>36.9</td>
<td>14.3</td>
<td>23.7</td>
<td>23.2</td>
<td>20.0</td>
</tr>
<tr>
<td>SWBD</td>
<td>9.0</td>
<td>30.6</td>
<td>6.1</td>
<td>4.6</td>
<td>11.1</td>
<td>11.4</td>
</tr>
<tr>
<td>ETT</td>
<td>18.1</td>
<td>35.3</td>
<td>17.4</td>
<td>22.1</td>
<td>10.3</td>
<td>16.6</td>
</tr>
<tr>
<td>WSJ</td>
<td>23.1</td>
<td>41.1</td>
<td>22.5</td>
<td>30.1</td>
<td>25.4</td>
<td>14.2</td>
</tr>
<tr>
<td>NANC</td>
<td>20.4</td>
<td>39.8</td>
<td>19.3</td>
<td>27.1</td>
<td>24.5</td>
<td>18.3</td>
</tr>
</tbody>
</table>
Model and estimation

\[
predict(\mathbf{X}, \mathbf{y}) = \lambda_c \cosinesim(\mathbf{X}, \mathbf{y}) + \lambda_w \text{unkwords}(\mathbf{X}, \mathbf{y}) + \lambda_u \text{entropy}(\mathbf{X}) + b
\]
Model and estimation

\[ \text{predict}(\vec{\lambda}_c, \text{cosinesim}(\vec{\lambda}_c, \text{unkwords}(\vec{\lambda}_w, \text{entropy}(\vec{\lambda}_u)) \ldots) = \]
Training data

* numbers on this slide are cooked
Training data

89% 88%
83% 84%
86% 83%
82% 87%
83% 85%
78% 81%

* numbers on this slide are cooked
### Corpora used

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Source domain</th>
<th>Target domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>BROWN</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ETT</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>GENIA</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>PUBMED</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>SWBD</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>WSJ</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NANC</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>GUTENBERG</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>
## Corpora used

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Source domain</th>
<th>Target domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>BROWN</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ETT</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>GENIA</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>PUBMED</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>SWBD</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>WSJ</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NANC</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>GUTENBERG</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>
### Corpora used

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Source domain</th>
<th>Target domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>BROWN</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ETT</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>GENIA</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>PUBMED</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>SWBD</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>WSJ</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NANC</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>GUTENBERG</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Round-robin evaluation
Round-robin evaluation
Evaluation for GENIA

**train sources**

- WSJ
- Brown
- ETT

**targets**

- WSJ
- Brown
- ETT

BNC
Evaluation for GENIA

**train sources**
- WSJ
- Brown
- ETT
- NANC
- Gutenberg

**targets**
- WSJ
- Brown
- ETT
- BNC

**test sources**
- WSJ
- Brown
- ETT
- NANC
- Gutenberg

**target**
- GENIA
Baselines

- **Standard baselines**
  - Uniform with labeled corpora
  - Uniform with labeled and self-trained corpora
  - Fixed set: WSJ

- **Oracle baselines**
  - Best single corpus
  - Best seen
Evaluation results
Evaluation results
Evaluation results

![Graph showing evaluation results with various data points representing different corpora and settings. The x-axis represents different corpus types, including 'Best single corpus', 'Fixed set: WSJ', 'Uniform', and 'Self-trained Uniform'. The y-axis represents f-score values from 74 to 87. Various markers indicate different corpora and settings, such as 'Brown', 'ETT', 'BNC', 'WSJ', 'Switchboard', and 'GENIA'.]
Evaluation results
Moral of the story

- Domain differences can be captured by surface features
- Any Domain Parsing:
  - near-optimal performance for out-of-domain evaluation
  - domain-specific parsing models are beneficial
- Self-trained corpora improve accuracy across domains
Future work

In order of decreasing $\frac{bang}{buck}$:

- Automatically adapting the reranker (and other non-linear models)
- Other parsing model combination strategies
- Applying to other tasks
- Non-linear regression
- Syntactic features
May The Force Be With You

Questions?

Thanks to the members of the Brown, Berkeley, and Stanford NLP groups for their feedback and support!

Brought to you by NSF grants LIS9720368 and IIS0095940 and DARPA GALE contract HR0011-06-2-0001
Extra slides
Sampling parsing models

**Goal:** parsing models with many different subsets of corpora

1. Sample $n = \#$ source domains from exponential distribution
2. Sample probabilities for $n$ corpora from $n$-simplex
3. Sample names for $n$ corpora

Repeat until “done”
Average oracle $f$-score

![Graph showing the Average oracle $f$-score as a function of the Number of mixed parsing model samples. The graph indicates a steady increase in $f$-score until approximately 500 samples, after which the $f$-score plateaus at around 87.5.](image-url)
Out-of-domain evaluation for GENIA

**train**

**sources**

- WSJ
- Brown
- ETT
- NANC
- Gutenberg

**targets**

- WSJ
- Brown
- ETT
- BNC

**test**

**sources**

- WSJ
- Brown
- ETT
- NANC
- Gutenberg

**target**

- GENIA
In-domain evaluation for GENIA

**train**

sources

- WSJ
- Brown
- ETT
- NANC
- Gutenberg
- SWITCHBOARD
- PubMed

**targets**

- WSJ
- Brown
- ETT
- NANC
- Gutenberg
- SWITCHBOARD
- BNC

**test**

sources

- WSJ
- Brown
- ETT
- NANC
- Gutenberg
- SWITCHBOARD
- PubMed

**target**

- GENIA
Tuning parameters

- We want to select regression model, features
- Evaluation is round-robin
- Tuning can be done with nested round-robins
  - hold out one target corpus entirely
  - round-robin on each remaining target corpus
- This results in 30 small tuning scenarios
Tuning metrics

- Three metrics to do model/feature selection:
  - These metrics are summed across all 30 tuning scenarios
  - Parallelized best-first search explored 6,000 settings
  - Our best setting performs well over all three metrics: cosine ($k=50$), unknown words ($target \rightarrow source$), entropy
Tuning metrics

- Three metrics to do model/feature selection:
  1. mean squared error:
     \[ \sum (\text{true} - \text{predicted})^2 \]

- These metrics are summed across all 30 tuning scenarios
- Parallelized best-first search explored 6,000 settings
- Our best setting performs well over all three metrics: cosine \((k=50)\), unknown words \((\text{target} \rightarrow \text{source})\), entropy
Tuning metrics

- Three metrics to do model/feature selection:
  1. mean squared error:
     \[ \sum (\text{true} - \text{predicted})^2 \]
  2. modified mean squared error:
     \[ \sum |\text{true} - \text{predicted}|^{1+\text{true}} \]

- These metrics are summed across all 30 tuning scenarios
- Parallelized best-first search explored 6,000 settings
- Our best setting performs well over all three metrics: cosine \((k=50)\), unknown words \((\text{target} \rightarrow \text{source})\), entropy
Tuning metrics

- Three metrics to do model/feature selection:
  1. mean squared error:
     \[ \sum (\text{true} - \text{predicted})^2 \]
  2. modified mean squared error:
     \[ \sum |\text{true} - \text{predicted}|^{1+\text{true}} \]
  3. oracle loss:
     \[ \max(\text{true}) - \text{evaluate}(\max(\text{predicted})) \]

- These metrics are summed across all 30 tuning scenarios
- Parallelized best-first search explored 6,000 settings
- Our best setting performs well over all three metrics: cosine \((k=50)\), unknown words \((\text{target} \rightarrow \text{source})\), entropy
Feature interactions

<table>
<thead>
<tr>
<th></th>
<th>Cosine ((k = 50))</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Unknown words</td>
</tr>
<tr>
<td>-</td>
<td>Relative entropy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unknown words</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Relative entropy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Relative entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Cosine ((k = 50))</td>
</tr>
<tr>
<td>-</td>
<td>Relative entropy</td>
</tr>
</tbody>
</table>
Out-of-domain evaluation results

<table>
<thead>
<tr>
<th>Best single corpus</th>
<th>Fixed set WSJ</th>
<th>Uniform</th>
<th>Self-trained Uniform</th>
<th>Any Domain Parsing</th>
<th>Best seen</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>80</td>
<td>81</td>
<td>Average</td>
<td>82</td>
<td>83</td>
<td>84</td>
</tr>
<tr>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>90</td>
</tr>
<tr>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>101</td>
</tr>
<tr>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
<td>106</td>
</tr>
</tbody>
</table>

Corpora: BNC, GENIA, Brown, Switchboard, ETT, WSJ
In-domain evaluation results

- Fixed set
- Uniform
- Best single corpus
- Self-trained
- Any Domain Parsing
- Best overall model
- Best seen

- WSJ
- GENIA
- Brown
- Switchboard
- ETT
- BNC

Average scores:
- BNC
- GENIA
- Brown
- Switchboard
- ETT
- WSJ