Advances in cross-lingual syntactic transfer

Ryan McDonald
O. Tackstrom, S. Petrov, K. Hall, J. Nivre
+ others
An Empirical Case Against Unsupervised Parsing

Ryan McDonald
O. Tackstrom, S. Petrov, K. Hall, J. Nivre
+ others
1. (Labeled) head-modifier relations
2. Part-of-speech tags

Used in: Translation, IE, Voice Actions, Sentiment
Syntactic Parsing

1. (Labeled) head-modifier relations
2. Part-of-speech tags

Used in: Translation, IE, Voice Actions, Sentiment
1. (Labeled) head-modifier relations
2. Part-of-speech tags

Used in: Translation, IE, Voice Actions, Sentiment
MST-pars**er**

- Discriminative CRF feature-rich parser

\[ P_\theta(y|x) = \frac{\exp\{\theta \cdot \Phi(y)\}}{Z(x)} \]

- Arc-factored model

\[ \Phi(y) = \sum_{h \rightarrow m \in y} \phi(h \rightarrow m) \]

- Trained with L-BFGS

\[ \theta = \arg \max_{\theta} \sum_{(x,y)} \log P_\theta(y|x) \]

McDonald et al. 05
**Goal:** Parsers that produce a common syntactic representation for all the world’s languages.
**Parsing The World’s Languages**

- **Goal**: Parsers that produce a common syntactic representation for all the world’s languages
- Treebanks for ~20 languages
  - Quality varies substantially across languages
  - Annotation schemes vary across languages
  - Some high profile languages not covered
**Goal**: Parsers that produce a common syntactic representation for all the world’s languages

- Treebanks for ~20 languages
  - Quality varies substantially across languages
  - Annotation schemes vary across languages
  - Some high profile languages not covered

- No corpora needed: Unsupervised parsing (DMV++)
  - Accuracy: unsupervised << supervised
  - Unlabeled, short sentences only, designed for En ...
  - Not practical for most applications
Syntactic Transfer

- Learn parsers for resource-poor languages from resource-rich languages
- Hwa et al. 2005 and earlier

English Treebank

- John
- likes
- Mary

NOUN
VERB
NOUN

Syntactic Transfer

- ...
Single Source Transfer

- Parallel data / annotation projection (Hwa et al. 2005)
- **Delexicalized direct transfer** (no parallel data)
  - Parse foreign language with cross-lingual features
  - E.g., part-of-speech tags, clusters, dictionaries
Delexicalized Transfer

**English TB**

- John (NOUN)
- likes (VERB)
- Mary (NOUN)

...
Delexicalized Transfer

English TB

Parser

Ignore words

John NOUN
likes VERB
Mary NOUN
Delexicalized Transfer

English TB

Delexicalized Parser

Ignore words
Delexicalized Transfer

English TB

Delexicalized Parser

NOUN VERB NOUN

...
Delexicalized Transfer

English TB

Delexicalized Parser

Ο Γιάννης βλέπει την Μαρία

DET  NOUN  VERB  DET  NOUN

NOUN  VERB  NOUN
Delexicalized Transfer

English TB

Delexicalized Parser

O Γιαννης βλεπει την Μαρια

DET NOUN VERB DET NOUN
Delexicalized Transfer

English TB

Delexicalized Parser

Zeman&Resnik ’08: POS-tags
Täckström et al. ‘12: xling clusters
Durrett et al. ‘12: xling dict
Delexicalized Transfer (En)

Arc-factored CRF/ML MST-parser
Delexicalized Transfer (En)

Arc-factored CRF/ML MST-parser

Indo-European
Arc-factored CRF/ML MST-parser

Delexicalized Transfer (En)

Non-Indo-European

Indo-European
Delexicalized Transfer (En)

Comparison with best weakly/unsupervised systems

- **EN-Delex**
- **Naseem et al. 2010**
- **EN-Delex**
- **Marecek & Zabokrtsky 2012**
- **Spitkovsky et al. 2012**

AVG-7-langs <=10 words:
- 71.1
- 56.9

AVG-16-langs:
- 52.1
- 41.1
- 46.7
Delexicalized Transfer (En)

Comparison with best weakly/unsupervised systems

AVG-7-langs <=10 words:
- EN-Delex: 71.1
- Naseem et al. 2010: 56.9

EN-Delex
EN-Delex
Marecek & Zabokrtsky 2012
Spitkovsky et al. 2012

AVG-16-langs:
- EN-Delex: 52.1
- Naseem et al. 2010: 41.1
- Spitkovsky et al. 2012: 46.7
Multi-Source Transfer

- Delexicalized mixtures
  - Parallel data less important/possible
    - Maybe only En<->XX available
  - McDonald et al. ‘11, Cohen et al. ‘11, Søgaard ‘11
Multi-Source Transfer

- Simple solution (McDonald et al ‘11):
  - Concatenate source language treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>En-Delex</th>
<th>Multi-source Delex</th>
</tr>
</thead>
<tbody>
<tr>
<td>ba</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>ar</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td>ja</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td>cz</td>
<td>41</td>
<td>37</td>
</tr>
<tr>
<td>tr</td>
<td>37</td>
<td>51</td>
</tr>
<tr>
<td>hu</td>
<td>56</td>
<td>52</td>
</tr>
<tr>
<td>nl</td>
<td>58</td>
<td>53</td>
</tr>
<tr>
<td>de</td>
<td>56</td>
<td>53</td>
</tr>
<tr>
<td>zh</td>
<td>57</td>
<td>53</td>
</tr>
<tr>
<td>el</td>
<td>59</td>
<td>63</td>
</tr>
<tr>
<td>bg</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>sv</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>es</td>
<td>67</td>
<td>64</td>
</tr>
<tr>
<td>it</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>pt</td>
<td>78</td>
<td>73</td>
</tr>
<tr>
<td>ca</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>AVG</td>
<td>55</td>
<td>55</td>
</tr>
</tbody>
</table>
Selective-Sharing
(Naseem et al. 2012)
Selective-Sharing
(Naseem et al. 2012)

- Generative model with two components
Selective-Sharing
(Naseem et al. 2012)

- Generative model with two components
  - Head-modifier preferences (language independent)
  - "Adjectives modify nouns"
  - Learned from all languages
Selective-Sharing
(Naseem et al. 2012)

- Generative model with two components
  - Head-modifier preferences (language independent)
    - “Adjectives modify nouns”
    - Learned from all languages
  - Ordering preferences (language dependent)
    - “Adjectives modify nouns to their right”
    - Selectively shares parameters between languages
    - Shares through typological properties from the World Atlas of Language, e.g., NOUN-ADJ vs ADJ-NOUN
Selective-Sharing
(Naseem et al. 2012)

MST-Delex
Naseem Selective-Sharing
Selective-Sharing
(Naseem et al. 2012)

- MST-Delex
- Naseem Selective-Sharing

Bar chart comparing MST-Delex and Naseem Selective-Sharing across various languages with an average (AVG) value.
Selective-Sharing

(NASEEM ET AL. 2012)

MST-Delex
Naseem Selective-Sharing

Non-Indo-European
Selective-Sharing
(Naseem et al. 2012)

MST-Delex
Naseem Selective-Sharing

Indo-European
Non-Indo-European

ba  tr  ja  cz  ar  zh  hu  de  nl  sv  el  bg  es  it  ca  pt  AVG
30  40  37  39  41  44  43  45  40  45  45  48  48  45  43  55  59

AVG
Selective-Sharing
(Naseem et al. 2012)

- Impoverished generative model
  - Poor for source-heavy Indo-European langs
  - Better supervised accuracies with discriminative feature rich models

<table>
<thead>
<tr>
<th>Language</th>
<th>Naseem et al.</th>
<th>MST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>64.2</td>
<td>82.6</td>
</tr>
<tr>
<td>Basque</td>
<td>51.6</td>
<td>68.7</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>71.0</td>
<td>90.0</td>
</tr>
<tr>
<td>Catalan</td>
<td>72.1</td>
<td>91.7</td>
</tr>
<tr>
<td>Chinese</td>
<td>73.5</td>
<td>83.4</td>
</tr>
<tr>
<td>Czech</td>
<td>58.9</td>
<td>86.6</td>
</tr>
<tr>
<td>Dutch</td>
<td>58.0</td>
<td>83.1</td>
</tr>
<tr>
<td>German</td>
<td>58.0</td>
<td>84.3</td>
</tr>
<tr>
<td>Greek</td>
<td>70.5</td>
<td>79.9</td>
</tr>
<tr>
<td>Hungarian</td>
<td>61.6</td>
<td>83.9</td>
</tr>
<tr>
<td>Italian</td>
<td>72.3</td>
<td>92.2</td>
</tr>
<tr>
<td>Japanese</td>
<td>75.6</td>
<td>76.3</td>
</tr>
<tr>
<td>Portuguese</td>
<td>79.6</td>
<td>88.3</td>
</tr>
<tr>
<td>Spanish</td>
<td>65.3</td>
<td>87.3</td>
</tr>
<tr>
<td>Swedish</td>
<td>73.0</td>
<td>78.4</td>
</tr>
<tr>
<td>Turkish</td>
<td>67.6</td>
<td>89.3</td>
</tr>
</tbody>
</table>

Average: 67.1 | 84.1
Selective sharing can be seen as a form of target language adaptation.

Natural Question:

* Can we adapt discriminative feature-rich models to specific target languages?

Two orthogonal approaches

1. Selective sharing for MST models
2. Re-lexicalization via ambiguity-preserving training
MST-parser Selective Sharing

- MST-Delex (55.1% -- average over 16 langs)
  - Full MST features, delexicalized
MST-parser Selective Sharing

- MST-Delex (55.1% -- average over 16 langs)
  - Full MST features, delexicalized
- MST-Bare (51.5%)
  - Reduced MST features; no direction or complex n-grams
  - Models attachment preferences only
MST-parser Selective Sharing

- MST-Delex (55.1% -- average over 16 langs)
  - Full MST features, delexicalized
- MST-Bare (51.5%)
  - Reduced MST features; no direction or complex n-grams
  - Models attachment preferences only
- MST-Share** (57.4%)
  - Features for specific tags + WALS properties
  - Conjoined with direction of attachment
## MST-parser Selective Sharing

- **MST-Delex** *(55.1% -- average over 16 langs)*
  - Full MST features, delexicalized
- **MST-Bare** *(51.5%)*
  - Reduced MST features; no direction or complex n-grams
  - Models attachment preferences only
- **MST-Share** *(57.4%)*
  - Features for specific tags + WALS properties
  - Conjoined with direction of attachment
- **MST-Family** *(62.0%)*
  - Add full MST features shared language family

**Selective Sharing**
MST-parser **Selective Sharing**

- MST-Delex (55.1% -- average over 16 langs)
  - Full MST features, delexicalized
- MST-Bare (51.5%)
  - Reduced MST features; no direction or complex n-grams
  - Models attachment preferences only
- MST-Share** (57.4%)
  - Features for specific tags + WALS properties
  - Conjoined with direction of attachment
- MST-Family** (62.0%)
  - Add full MST features shared language family
- MST-Bare ⊂ MST-Share ⊂ MST-Family
MST-Share:

\[ \text{dir} \otimes \text{wals.85A} \otimes \text{head.tag=ADP} \otimes \text{modifier.tag=NOUN} \]

order of adposition

MST-Family:

\[ f \otimes \text{lang-family}, \forall f \in \text{MST-delex} \]

\[ f = \text{Every full MST feature} \]
\[ 2\text{-tier lang-family} = \{\text{Indo-European}\}, \{\text{Altaic}\}, +\text{Isolates} \]
1. 7% absolute improvement
2. 15% relative reduction in errors
3. Ar, Eu, Hu, Ja, Tr & Zh
4. 7% relative reduction in errors over Naseem et al.
MST-parsen Selective Sharing

- MST-Delex
- Naseem Selective-Sharing
- MST-Selective-Sharing

Bar chart showing performance metrics for different languages.
MST-parser Selective Sharing

- MST-Delex
- Naseem Selective-Sharing
- MST-Selective-Sharing

AVG
MST-parser Selective Sharing

- MST-Delex
- Naseem Selective-Sharing
- MST-Selective-Sharing

Non-Indo-European
# Oracle Observations (i)

<table>
<thead>
<tr>
<th>MST Selective-Sharing</th>
<th>MST-supervised Delexicalized</th>
<th>MST-supervised Lexicalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.0</td>
<td>78.9</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Lexical Features Help
<table>
<thead>
<tr>
<th></th>
<th>MST Selective-Sharing</th>
<th>MST-supervised Delexicalized</th>
<th>MST-supervised Lexicalized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62.0</td>
<td>78.9</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Lexical Features Help

Target Language Re-lexicalization with Self-Training

Zeman & Resnik '08
## Oracle Observations (1)

<table>
<thead>
<tr>
<th></th>
<th>MST Selective-Sharing</th>
<th>MST-supervised Delexicalized</th>
<th>MST-supervised Lexicalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.0</td>
<td>78.9</td>
<td>84.1</td>
<td></td>
</tr>
</tbody>
</table>

Lexical Features Help

Target Language Re-lexicalization with Self-Training

62.0 → 62.4

Zeman & Resnik ’08
## Oracle Observations (II)

<table>
<thead>
<tr>
<th></th>
<th>MST</th>
<th>MST-supervised Delexicalized</th>
<th>MST-supervised Lexicalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.0</td>
<td>78.9</td>
<td>84.1</td>
<td></td>
</tr>
</tbody>
</table>

A lot of good ambiguity at top of prediction space

| k-best Oracle MST Selective-Sharing | 73.2 |
Ambiguity-Preserving Self-Training

Indo-European -> Japanese

100x
a v
N V

a b v
N N V

a b v
N N V
Ambiguity-Preserving Self-Training

Indo-European -> Japanese

100x

100%

a v
N V

60%  100%

a b v
N N V

60%  100%

a b v
N N V
Ambiguity-Preserving Self-Training

Indo-European -> Japanese

100x

a v
N V

60%
100%

a b v
N N V

40%

a b v
N N V

40%

100%
Indo-European -> Japanese

Allowing the 40% edges during training gives the feature-rich model room to learn lexical preferences: v -> a
Ambiguity-Preserving Self-Training

- Regular self-training reinforces errors
- Want to capture base model’s uncertainty
- Allow self-trainer freedom to back-off from Viterbi parse
Ambiguity-Preserving Self-Training

- Regular self-training reinforces errors
- Want to capture base model’s uncertainty
- Allow self-trainer freedom to back-off from Viterbi parse

\( \tilde{\mathcal{Y}}(x) \) : space of self-training trees, regular self-training = Viterbi parse
Ambiguity-Preserving Self-Training

- Regular self-training reinforces errors
- Want to capture base model’s uncertainty
- Allow self-trainer freedom to back-off from Viterbi parse

\[ \tilde{\mathcal{Y}}(x) : \text{space of self-training trees}, \]
\[ \text{regular self-training} = \text{Viterbi parse} \]

We want to give learner flexibility to move probability mass to any of the trees in this set.
Ambiguity-Preserving Self-Training

- Regular self-training reinforces errors
- Want to capture base model’s uncertainty
- Allow self-trainer freedom to back-off from Viterbi parse

\[ \tilde{Y}(x) : \text{space of self-training trees,} \]
\[ \text{regular self-training} = \text{Viterbi parse} \]

We want to give learner flexibility to move probability mass to any of the trees in this set.

Needs to be large enough to contain good trees, but small enough to guide learner.
1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)
Ambiguity-Preserving Self-Training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)
1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

3. Let \(\tilde{\mathcal{Y}}(x)\) be the set of trees derivable from \(A(x, \ast)\)
Ambiguity-Preserving Self-Training

1. Sort arcs \((h \to m)\) by marginal prob: \(\mu(h \to m; x)\)

2. Add \((h \to m)\) to \(A(x, m)\) until \(\sum_{h \to m \in A(x, m)} \mu(h \to m; x) \geq \sigma\)

3. Let \(\tilde{\mathcal{Y}}(x)\) be the set of trees derivable from \(A(x, \ast)\)

4. Optimize: 
\[
\left( \sum_x \log \sum_{y \in \tilde{\mathcal{Y}}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|	heta\|^2
\]
Ambiguity-Preserving Self-Training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

3. Let \(\tilde{Y}(x)\) be the set of trees derivable from \(A(x, \ast)\)

4. Optimize:
\[
\left( \sum_{x} \log \sum_{y \in \tilde{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|\theta\|^2
\]

Regular self-training, this is just Viterbi parse
Ambiguity-Preserving Self-Training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

3. Let \(\hat{Y}(x)\) be the set of trees derivable from \(A(x, *)\)

4. Optimize: 
\[
\left( \sum_{x} \log \sum_{y \in \hat{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|\theta\|^2
\]

Same objective as multi/partial-label CRFs Riezler et al. 02

Regular self-training, this is just Viterbi parse
Ambiguity-Preserving Self-Training

1. 9% absolute improvement over base MST model
2. 4% absolute over Naseem+EM; better on 14/16 langs
3. 2% over MST selective sharing
4. Consistent across all languages
Ambiguity-Preserving Ensemble-Training

Original Ambiguity-Preserving Self-training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

3. Let \(\tilde{Y}(x)\) be the set of trees derivable from \(A(x, \ast)\)

4. Optimize: \(\left( \sum_{x} \log \sum_{y \in \tilde{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|	heta\|^2\)
Ambiguity-Preserving Ensemble-Training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

3. Let \(\tilde{Y}(x)\) be the set of trees derivable from \(A(x, *)\)

4. Optimize:
\[
\left( \sum_x \log \sum_{y \in \tilde{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|\theta\|^2
\]
Ambiguity-Preserving Ensemble-Training

1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x, m)} \mu(h \rightarrow m; x) \geq \sigma\)

2b. Let there be \(k\) such \(A_k(x, m)\) from \(k\) parsers

3. Let \(\tilde{Y}(x)\) be the set of trees derivable from \(A(x, *)\)

4. Optimize: \(\left( \sum_x \log \sum_{y \in \tilde{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \|\theta\|^2\)
1. Sort arcs \((h \rightarrow m)\) by marginal prob: \(\mu(h \rightarrow m; x)\)

2. Add \((h \rightarrow m)\) to \(A(x, m)\) until \(\sum_{h \rightarrow m \in A(x,m)} \mu(h \rightarrow m; x) \geq \sigma\)

2b. Let there be \(k\) such \(A_k(x, m)\) from \(k\) parsers

2c. Let \(A(x, m) = \bigcup_k A_k(x, m)\)

3. Let \(\tilde{Y}(x)\) be the set of trees derivable from \(A(x, \ast)\)

4. Optimize: \(\left( \sum_x \log \sum_{y \in \tilde{Y}(x)} P(y|x; \theta) \right) + \frac{\lambda}{2} \| \theta \|^2\)
Ambiguity-Preserving Ensemble-Training

1. 10.3% absolute improvement over base MST model
2. 5% absolute over Naseem +EM, better on 15/16 langs
3. 3.4% over MST selective sharing
4. Consistent across all languages

* base parsers

MST Selective Sharing +APET

AVG-16-langs
1. 10.3% absolute improvement over base MST model
2. 5% absolute over Naseem + EM, better on 15/16 langs
3. 3.4% over MST selective sharing
4. Consistent across all languages

SOTA Unsupervised
Spitkovsky et al. 2012
16 lang avg: **46.7**
Final Thoughts

- Since 2010 (Naseem et al., delexicalized transfer, selective sharing)
  - 20% absolute improvements over SOTA unsupervised parsers
  - Under estimate (IMO) due to treebank divergences
  - All sentence lengths, no magic initializations, ...
  - Simple models + feature engineering
Final Thoughts

- Since 2010 (Naseem et al., delexicalized transfer, selective sharing)
  - 20% absolute improvements over SOTA unsupervised parsers
    - Under estimate (IMO) due to treebank divergences
    - All sentence lengths, no magic initializations, ...
    - Simple models + feature engineering
- Unsupervised parsing??
  - From engineering perspective = not really needed
  - From psycholinguistic perspective = ??
Final Thoughts

- Since 2010 (Naseem et al., delexicalized transfer, selective sharing)
  - 20% absolute improvements over SOTA unsupervised parsers
  - Under estimate (IMO) due to treebank divergences
  - All sentence lengths, no magic initializations, ...
  - Simple models + feature engineering
- Unsupervised parsing??
  - From engineering perspective = not really needed
  - From psycholinguistic perspective = ??
- Next steps: Language specific constraints (via PR?)
Final Thoughts

- Since 2010 (Naseem et al., delexicalized transfer, selective sharing)
  - 20% absolute improvements over SOTA unsupervised parsers
  - Under estimate (IMO) due to treebank divergences
  - All sentence lengths, no magic initializations, ...
  - Simple models + feature engineering
- Unsupervised parsing??
  - From engineering perspective = not really needed
  - From psycholinguistic perspective = ??
- Next steps: Language specific constraints (via PR?)
- Next steps: Better “universal representations”
Final Thoughts

- Since 2010 (Naseem et al., delexicalized transfer, selective sharing)
  - 20% absolute improvements over SOTA unsupervised parsers
  - Under estimate (IMO) due to treebank divergences
  - All sentence lengths, no magic initializations, ...
  - Simple models + feature engineering
- Unsupervised parsing??
  - From engineering perspective = not really needed
  - From psycholinguistic perspective = ??
- Next steps: Language specific constraints (via PR?)
- Next steps: Better “universal representations”
- Next steps: Better evaluation with universal treebanks
Thanks!

- Collaborators: Oscar Täckström, Slav Petrov, Keith Hall, Joakim Nivre
- Google parsing team: Dipanjan Das, Yoav Goldberg, Hao Zhang, Kuzman Ganchev, Michael Ringaard, Alexander Rush
- Others: Shay Cohen, Noah Smith, Anders Søgaard, Tahira Naseem and Regina Barzilay