Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance

Shay B. Cohen      Dipanjan Das      Noah A. Smith

Carnegie Mellon University

July 27
EMNLP 2011
Goal:

Learn linguistic structure for a language without any labeled data in that language

Cohen, Das and Smith (2011)
Goal:

Learn linguistic structure for a language without any labeled data in that language

The Skibo Castle is close by.
Goal:

Learn linguistic structure for a language without any labeled data in that language

The Skibo Castle is close by.

Part-of-Speech Tagging

Cohen, Das and Smith (2011)
Goal:
Learn linguistic structure for a language without any labeled data in that language

The Skibo Castle is close by.

Dependency Parsing

Cohen, Das and Smith (2011)
Multilingual Unsupervised Learning
Multilingual Unsupervised Learning

- using parallel data
- no parallel data

Cohen, Das and Smith (2011)
Multilingual Unsupervised Learning

using parallel data

no parallel data (hard)
Multilingual Unsupervised Learning

- using parallel data
  - joint learning for multiple languages
    - Snyder et al. (2009)
    - Naseem et al. (2010)
  - supervision in source language(s)
    - Yarowsky and Ngai (2001)
    - Xi and Hwa (2005)
    - Smith and Eisner (2009)
    - Das and Petrov (2011)
    - McDonald et al. (2011)
- no parallel data (hard)

Cohen, Das and Smith (2011)
Multilingual Unsupervised Learning

using parallel data

no parallel data (hard)

joint learning for multiple languages
Snyder et al. (2009)
Naseem et al. (2010)

supervision in source language(s)
Yarowsky and Ngai (2001)
Xi and Hwa (2005)
Smith and Eisner (2009)
Das and Petrov (2011)
McDonald et al. (2011)

joint learning for multiple languages
Cohen and Smith (2009)
Multilingual Unsupervised Learning

- Using parallel data:
  - Joint learning for multiple languages
    - Snyder et al. (2009)
    - Naseem et al. (2010)
  - Supervision in source language(s)
    - Yarowsky and Ngai (2001)
    - Xi and Hwa (2005)
    - Smith and Eisner (2009)
    - Das and Petrov (2011)
    - McDonald et al. (2011)

- No parallel data (hard):
  - Joint learning for multiple languages
    - Cohen and Smith (2009)
  - Supervision in source language(s)
Multilingual Unsupervised Learning

- using parallel data
  - joint learning for multiple languages
    - Snyder et al. (2009)
    - Naseem et al. (2010)
  - supervision in source language(s)
    - Yarowsky and Ngai (2001)
    - Xi and Hwa (2005)
    - Smith and Eisner (2009)
    - Das and Petrov (2011)
    - McDonald et al. (2011)
- no parallel data (hard)
  - joint learning for multiple languages
    - Cohen and Smith (2009)
  - supervision in source language(s)

This work!

Cohen, Das and Smith (2011)
In a Nutshell

Annotated data + Unlabeled data in Portuguese = Portuguese parameters

Spanish | Italian

Cohen, Das and Smith (2011)  EMNLP 2011
In a Nutshell

**Annotated data** + **Unlabeled data in Portuguese** = **Portuguese parameters**

Coarse, *universal* parameters

Cohen, Das and Smith (2011)
In a Nutshell

Annotated data

Unlabeled data in Portuguese

= Portuguese parameters

Spanish

Italian

Coarse, universal parameters

Coarse, universal parameters

Interpolation (unsupervised training)

course parameters of Portuguese

Cohen, Das and Smith (2011)
In a Nutshell

Annotated data
Unlabeled data
= Portuguese parameters

Spanish
Italian

Monolingual unsupervised training in Portuguese

Coarse-to-fine expansion and initialization

course parameters of Portuguese

Cohen, Das and Smith (2011)
In a Nutshell

\[ \text{Annotated data} + \text{Unlabeled data in Portuguese} = \text{Portuguese parameters} \]

Monolingual unsupervised training in Portuguese

Cohen, Das and Smith (2011)
Assumptions for a given problem:

1. Underlying model is generative
Assumptions for a given problem:

1. Underlying model is generative

HMM

Merialdo (1994)
Assumptions for a given problem:

1. Underlying model is generative

DMV

Klein and Manning (2004)
Assumptions for a given problem:

1. Underlying model is generative

Composed of multinominal distributions
Assumptions for a given problem:

I. Underlying model is generative

Composed of multinomial distributions

HMM

The Skibo Castle is close by

Merialdo (1994)
Assumptions for a given problem:

1. Underlying model is generative

Composed of multinomial distributions

The Skibo Castle is close by

HMM

Merialdo (1994)
Assumptions for a given problem:

1. Underlying model is generative

Composed of multinomial distributions

\[ \theta_c \xleftarrow{p} \]

DMV

Klein and Manning (2004)
Assumptions for a given problem:

1. Underlying model is generative

In general, unlexicalized parameters $\theta$ look like:

$$\theta_{k, i}$$
Assumptions for a given problem:

1. Underlying model is generative

In general, unlexicalized parameters $\theta$ look like:

$$\theta_{k,i}$$

$k^{th}$ multinomial in the model
Assumptions for a given problem:

I. Underlying model is generative

In general, unlexicalized parameters $\theta$ look like:

$$\theta_{k,i}$$

$k^{th}$ multinomial in the model \hspace{2cm} $i^{th}$ event in the multinomial
Assumptions for a given problem:

1. Underlying model is generative

In general, unlexicalized parameters $\theta$ look like:

$$\theta_{k,i}$$

$k^{th}$ multinomial in the model

$i^{th}$ event in the multinomial

e.g. transition from ADJ $(k)$ to NOUN $(i)$

Cohen, Das and Smith (2011)
Assumptions for a given problem:

1. Underlying model is generative

The lexicalized parameters $\eta$ take a similar form
(No lexicalized parameters for the DMV)
Assumptions for a given problem:

1. Underlying model is generative

\[ p(\text{sentence}, \text{derivation}) = \]
Assumptions for a given problem:

1. Underlying model is generative

\[ p(\text{sentence, derivation}) = \prod_{k=1}^{K} \prod_{i=1}^{N_k} \theta^{f_{k,i}}_{k,i} \prod_{m=1}^{M} \prod_{i=1}^{N_m} \eta^{f_{m,i}}_{m,i} \]

unlexicalized

lexicalized

Cohen, Das and Smith (2011)
Assumptions for a given problem:

1. Underlying model is generative

\[
p(\text{sentence, derivation}) = \prod_{k=1}^{K} \prod_{i=1}^{N_k} \theta_{k,i}^{f_{k,i}} \prod_{m=1}^{M} \prod_{i=1}^{N_m} \eta_{m,i}^{f_{m,i}}
\]

number of times event \( i \) of multinomial \( k \) fires in the derivation

unlexicalized \hspace{1cm} \text{lexicalized}
Assumptions for a given problem:

2. Coarse, universal part-of-speech tags
Assumptions for a given problem:

2. Coarse, universal part-of-speech tags

Petrov, Das and McDonald (2011)

VERB       DET
NOUN       CONJ
PRON       NUM
ADJ        PRT
ADV        X
ADP
Assumptions for a given problem:

2. Coarse, universal part-of-speech tags

For each language \( \ell \), there is a mapping

Treebank tagset

VERB  DET
NOUN  CONJ
PRON  NUM
ADJ  PRT
ADV  .
ADP  X

Cohen, Das and Smith (2011)
Assumptions for a given problem:

3. $L$ helper languages
Assumptions for a given problem:

3. $L$ helper languages

For each:

Treebank \rightarrow \text{coarse conversion} \rightarrow \text{Coarse treebank}
Assumptions for a given problem:

3. \( L \) helper languages

For each:

Treebank \rightarrow \text{coarse conversion} \rightarrow \text{Coarse treebank} \rightarrow \text{MLE} \rightarrow \text{unlexicalized parameters} \( \theta^{(L)} \)

Cohen, Das and Smith (2011)
Multilingual Modeling
Multilingual Modeling

For a target language, unlexicalized parameters:
Multilingual Modeling

For a target language, unlexicalized parameters:

\[ \theta_k = \sum_{\ell=1}^{L} \beta_{\ell,k} \theta_k^{(\ell)} \]
For a target language, unlexicalized parameters:

\[ \theta_k = \sum_{\ell=1}^{L} \beta_{\ell,k} \theta_k^{(\ell)} \]

\( k^{th} \) multinomial in the model
(say, the transitions from the ADJ tag in an HMM)
Multilingual Modeling

For a target language, unlexicalized parameters:

$$
\theta_k = \sum_{\ell=1}^{L} \beta_{\ell, k} \theta_k^{(\ell)}
$$

- $k^{th}$ multinomial in the model (say, the transitions from the ADJ tag in an HMM)
- Mixture weight for $k^{th}$ multinomial for the $\ell^{th}$ helper language

Cohen, Das and Smith (2011)

EMNLP 2011
Multilingual Modeling

e.g., two helper languages: Spanish and Italian
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

\[ \theta_{\text{ADJ} \rightarrow .} = \]
Multilingual Modeling

\[ \theta_{\text{ADJ} \rightarrow .} = \beta_{\text{Spanish, ADJ} \rightarrow .} \cdot \theta^{\text{Spanish}} + \beta_{\text{Italian, ADJ} \rightarrow .} \cdot \theta^{\text{Italian}} \]

e.g., two helper languages: Spanish and Italian

Cohen, Das and Smith (2011)
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

$$\theta_{\text{ADJ} \rightarrow .} = \beta_{\text{Spanish, ADJ} \rightarrow .} \cdot \theta_{\text{ADJ} \rightarrow .}^{(\text{Spanish})} + \beta_{\text{Italian, ADJ} \rightarrow .} \cdot \theta_{\text{ADJ} \rightarrow .}^{(\text{Italian})}$$

<table>
<thead>
<tr>
<th>NOUN</th>
<th>0.27</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERB</td>
<td>0.12</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.03</td>
</tr>
<tr>
<td>ADV</td>
<td>0.04</td>
</tr>
<tr>
<td>PRON</td>
<td>0.04</td>
</tr>
<tr>
<td>DET</td>
<td>0.03</td>
</tr>
<tr>
<td>ADP</td>
<td>0.25</td>
</tr>
<tr>
<td>NUM</td>
<td>0.01</td>
</tr>
<tr>
<td>CONJ</td>
<td>0.10</td>
</tr>
<tr>
<td>PRT</td>
<td>0.05</td>
</tr>
<tr>
<td>.</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

\[ \theta_{\text{ADJ} \rightarrow .} = \beta_{\text{Spanish, ADJ} \rightarrow .} \cdot \theta_{(\text{Spanish})} + \beta_{\text{Italian, ADJ} \rightarrow .} \cdot \theta_{(\text{Italian})} \]

Cohen, Das and Smith (2011)
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

\[ \theta_{\text{ADJ} \rightarrow .} = \beta_{\text{Spanish}, \text{ADJ} \rightarrow .} \cdot \theta_{(\text{Spanish})} + \beta_{\text{Italian}, \text{ADJ} \rightarrow .} \cdot \theta_{(\text{Italian})} \]
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

\[ \beta_{\text{Spanish}, \text{ADJ}} \cdot \theta_{\text{ADJ}} + \beta_{\text{Italian}, \text{ADJ}} \cdot \theta_{\text{ADJ}} \]

\begin{align*}
\text{NOUN} & \quad 0.27 \\
\text{VERB} & \quad 0.12 \\
\text{ADJ} & \quad 0.03 \\
\text{ADV} & \quad 0.04 \\
\text{PRON} & \quad 0.04 \\
\text{DET} & \quad 0.03 \\
\text{ADP} & \quad 0.25 \\
\text{NUM} & \quad 0.01 \\
\text{CONJ} & \quad 0.10 \\
\text{PRT} & \quad 0.05 \\
\text{.} & \quad 0.01 \\
\text{X} & \quad 0.05 \\
\end{align*}

\begin{align*}
\text{NOUN} & \quad 0.25 \\
\text{VERB} & \quad 0.11 \\
\text{ADJ} & \quad 0.04 \\
\text{ADV} & \quad 0.04 \\
\text{PRON} & \quad 0.06 \\
\text{DET} & \quad 0.04 \\
\text{ADP} & \quad 0.26 \\
\text{NUM} & \quad 0.0 \\
\text{CONJ} & \quad 0.20 \\
\text{PRT} & \quad 0.0 \\
\text{.} & \quad 0.0 \\
\text{X} & \quad 0.00 \\
\end{align*}
Learning and Inference
\[
\max_{\theta, \eta} \sum_{\text{sentences}} \log \sum_{\text{derivations}} p(\text{sentence, derivation})
\]

normal learning
Learning and Inference

\[
\max_{\theta, \eta} \sum_{\text{sentences}} \log \sum_{\text{derivations}} p(\text{sentence, derivation})
\]

**normal learning**
Learning and Inference

\[
\max_{\beta, \eta} \sum_{\text{sentences}} \log \sum_{\text{derivations}} p(\text{sentence, derivation})
\]

**multilingual learning**
Learning and Inference

\[ \max_{\boldsymbol{\beta}, \eta} \sum_{\text{sentences}} \log \sum_{\text{derivations}} p(\text{sentence, derivation}) \]

multilingual learning

\[ \theta_{k,i} = \sum_{\ell=1}^{L} \beta_{\ell,k} \theta_{k,i}^{(\ell)} \]
Learning and Inference

\[
\max_{\beta, \eta} \sum_{\text{sentences}} \log \sum_{\text{derivations}} p(\text{sentence, derivation})
\]

multilingual learning

\[
\theta_{k,i} = \sum_{\ell=1}^{L} \beta_{\ell,k} \theta_{k,i}^{(\ell)}
\]

\[
\theta^{(\ell)} \text{ are fixed!}
\]

Cohen, Das and Smith (2011)
Learning and Inference

Multilingual learning

learning $\beta$ with EM:
Learning and Inference

Multilingual learning

learning $\beta$ with EM:

$$E[g_{\ell,k}^{(t)}] = \sum_{\text{derivations}} p(\text{sentence, derivation}) \cdot g_{\ell,k}^{(t)}$$

Cohen, Das and Smith (2011)
Learning and Inference

Multilingual learning

Learning $\beta$ with EM:

$$E[g_{\ell,k}^{(t)}] = \sum_{\text{derivations}} p(\text{sentence, derivation}) \cdot g_{\ell,k}^{(t)}$$

Number of times $\beta_{\ell,k}$ is used in a derivation

Cohen, Das and Smith (2011)
Learning and Inference

**Multilingual learning**

**learning $\beta$ with EM:**

$$\mathbb{E}[g^{(t)}_{\ell,k}] = \sum_{\text{derivations}} p(\text{sentence, derivation}) \cdot g^{(t)}_{\ell,k}$$

**M-step:**

$$\beta_{\ell,k}^{(t+1)} = \frac{\mathbb{E}[g^{(t)}_{\ell,k}]}{\sum_{\ell'} \mathbb{E}[g^{(t)}_{\ell',k}]}$$
Learning and Inference

Multilingual learning

What about feature-rich generative models?
Learning and Inference

**Multilingual learning**

What about feature-rich generative models?

\[
\theta_{k,i} = \frac{\exp \psi^\top h(k,i)}{\sum_{i'} \exp \psi^\top h(k,i')}
\]

Berg-Kirkpatrick et al. (2010)
Learning and Inference

Multilingual learning

What about feature-rich generative models?

$$\theta_{k,i} = \frac{\exp \psi^\top h(k,i)}{\sum_{i'} \exp \psi^\top h(k,i')}$$

Locally normalized log-linear model

Berg-Kirkpatrick et al. (2010)
**Multilingual Modeling**

**e.g., two helper languages: Spanish and Italian**

\[
\beta_{\text{Spanish}, \text{ADJ}} \cdot \theta_{\text{ADJ}} + \beta_{\text{Italian}, \text{ADJ}} \cdot \theta_{\text{ADJ}}
\]

Cohen, Das and Smith (2011)
Multilingual Modeling

e.g., two helper languages: Spanish and Italian

\[
\beta_{\text{Spanish}, \text{ADJ} \rightarrow \cdot } \cdot \theta_{\text{Spanis}} + \beta_{\text{Italian, ADJ} \rightarrow \cdot } \cdot \theta_{\text{Italian}}
\]

Cohen, Das and Smith (2011)
**Multilingual Modeling**

e.g., two helper languages: Spanish and Italian

\[ \theta_{ADJ \rightarrow .} = \beta_{\text{Spanish}, \text{ADJ} \rightarrow .} \cdot \theta_{\text{(Spanish)} \rightarrow .} + \beta_{\text{Italian}, \text{ADJ} \rightarrow .} \cdot \theta_{\text{(Italian)} \rightarrow .} \]

<table>
<thead>
<tr>
<th>NOUN</th>
<th>0.26</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERB</td>
<td>0.12</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.03</td>
</tr>
<tr>
<td>ADV</td>
<td>0.04</td>
</tr>
<tr>
<td>PRON</td>
<td>0.05</td>
</tr>
<tr>
<td>DET</td>
<td>0.03</td>
</tr>
<tr>
<td>ADP</td>
<td>0.25</td>
</tr>
<tr>
<td>NUM</td>
<td>0.01</td>
</tr>
<tr>
<td>CONJ</td>
<td>0.13</td>
</tr>
<tr>
<td>PRT</td>
<td>0.04</td>
</tr>
<tr>
<td>.</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.04</td>
</tr>
</tbody>
</table>

\[0.6237 \cdot \theta_{\text{learned}} + 0.3763\]
Learning and Inference

Coarse-to-fine expansion

(for English)

\[ \theta_{\text{ADJ} \rightarrow .} \]

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>0.26</td>
</tr>
<tr>
<td>VERB</td>
<td>0.12</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.03</td>
</tr>
<tr>
<td>ADV</td>
<td>0.04</td>
</tr>
<tr>
<td>PRON</td>
<td>0.05</td>
</tr>
<tr>
<td>DET</td>
<td>0.03</td>
</tr>
<tr>
<td>ADP</td>
<td>0.25</td>
</tr>
<tr>
<td>NUM</td>
<td>0.01</td>
</tr>
<tr>
<td>CONJ</td>
<td>0.13</td>
</tr>
<tr>
<td>PRT</td>
<td>0.04</td>
</tr>
<tr>
<td>.</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)

EMNLP 2011
Learning and Inference

Coarse-to-fine expansion
(for English)

$\theta_{\text{ADJ}} \rightarrow \cdot$

$\theta_{\text{JJR}} \rightarrow \cdot$

$\theta_{\text{JJ}} \rightarrow \cdot$

$\theta_{\text{JJS}} \rightarrow \cdot$

<table>
<thead>
<tr>
<th></th>
<th>NOUN</th>
<th>VERB</th>
<th>ADJ</th>
<th>ADV</th>
<th>PRON</th>
<th>DET</th>
<th>ADP</th>
<th>NUM</th>
<th>CONJ</th>
<th>PRT</th>
<th>.</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.26</td>
<td>0.12</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>

identical copies

Step 1

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

\( \theta_{JJ} \rightarrow . \)

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>0.26</td>
</tr>
<tr>
<td>VERB</td>
<td>0.12</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.03</td>
</tr>
<tr>
<td>ADV</td>
<td>0.04</td>
</tr>
<tr>
<td>PRON</td>
<td>0.05</td>
</tr>
<tr>
<td>DET</td>
<td>0.03</td>
</tr>
<tr>
<td>ADP</td>
<td>0.25</td>
</tr>
<tr>
<td>NUM</td>
<td>0.01</td>
</tr>
<tr>
<td>CONJ</td>
<td>0.13</td>
</tr>
<tr>
<td>PRT</td>
<td>0.04</td>
</tr>
<tr>
<td>.</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion

(for English)

\[ \theta_{JJ} \rightarrow . \]

<table>
<thead>
<tr>
<th>NOUN</th>
<th>0.26</th>
</tr>
</thead>
<tbody>
<tr>
<td>VERB</td>
<td>0.12</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.03</td>
</tr>
<tr>
<td>ADV</td>
<td>0.04</td>
</tr>
<tr>
<td>PRON</td>
<td>0.05</td>
</tr>
<tr>
<td>DET</td>
<td>0.03</td>
</tr>
<tr>
<td>ADP</td>
<td>0.25</td>
</tr>
<tr>
<td>NUM</td>
<td>0.01</td>
</tr>
<tr>
<td>CONJ</td>
<td>0.13</td>
</tr>
<tr>
<td>PRT</td>
<td>0.04</td>
</tr>
<tr>
<td>.</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

Step 2

\[ \theta_{JJ} \rightarrow . \]

\[ \theta_{JJ} \rightarrow . \]

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

$\theta_{JJ} \rightarrow .$

Step 2

$\theta_{JJ} \rightarrow .$

Equal division

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

\[ \theta_{JJ} \rightarrow . \]

Step 2

\[ \theta_{JJ} \rightarrow . \]

Equal division

\[
\begin{array}{c|c|c|c|c|c|c|c}
\text{NOUN} & 0.26 &  &  &  &  &  & \\
\text{VERB} & 0.12 &  &  &  &  &  & \\
\text{ADJ} & 0.03 &  &  &  &  &  & \\
\text{ADV} & 0.04 &  &  &  &  &  & \\
\text{PRON} & 0.05 &  &  &  &  &  & \\
\text{DET} & 0.03 &  &  &  &  &  & \\
\text{ADP} & 0.25 &  &  &  &  &  & \\
\text{NUM} & 0.01 &  &  &  &  &  & \\
\text{CONJ} & 0.13 &  &  &  &  &  & \\
\text{PRT} & 0.04 &  &  &  &  &  & \\
. & 0.01 &  &  &  &  &  & \\
X & 0.04 &  &  &  &  &  & \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c|c|c|c}
\text{NN} & 0.065 &  &  &  &  &  & \\
\text{NNS} & 0.065 &  &  &  &  &  & \\
\text{NNP} & 0.065 &  &  &  &  &  & \\
\text{NNPS} & 0.065 &  &  &  &  &  & \\
\text{VB} & 0.02 &  &  &  &  &  & \\
\text{VBD} & 0.02 &  &  &  &  &  & \\
\text{VBG} & 0.02 &  &  &  &  &  & \\
\text{VBN} & 0.02 &  &  &  &  &  & \\
\text{VBP} & 0.02 &  &  &  &  &  & \\
\text{VBZ} & 0.02 &  &  &  &  &  & \\
\end{array}
\]

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

\[ \theta_{JJ} \rightarrow . \]

Step 2

Equal division

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

\[ \theta_{JJ} \rightarrow . \]

Monolingual unsupervised training

Initializer

Cohen, Das and Smith (2011)
Learning and Inference

Coarse-to-fine expansion
(for English)

\[ \theta_{JJ} \rightarrow . \]

Monolingual unsupervised training

New, fine \( \theta \)

Cohen, Das and Smith (2011)
Experiments
Two Problems
Two Problems

Unsupervised Part-of-Speech Tagging

Model: feature-based HMM
(Berg-Kirkpatrick et al., 2010)

Learning: L-BFGS

Cohen, Das and Smith (2011)
Two Problems

Unsupervised Part-of-Speech Tagging

Model:
feature-based HMM
(Berg-Kirkpatrick et al., 2010)

Learning:
L-BFGS

Unsupervised Dependency Parsing

Model:
DMV
(Klein and Manning, 2004)

Learning:
EM

Cohen, Das and Smith (2011)
Languages

Target Languages:

Bulgarian, Danish, Dutch, Greek, Japanese, Portuguese, Slovene, Spanish, Swedish, and Turkish

Helper Languages:

English, German, Italian and Czech

(CoNLL Treebanks from 2006 and 2007)
## Results: POS Tagging

<table>
<thead>
<tr>
<th></th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(without tag dictionary)

Cohen, Das and Smith (2011)
## Results: POS Tagging

<table>
<thead>
<tr>
<th></th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monolingual baseline (Berg-Kirkpatrick et al., 2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(without tag dictionary)

Cohen, Das and Smith (2011)
Results: POS Tagging

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
</table>

Uniform mixture parameters (no learning)

(without tag dictionary)

Cohen, Das and Smith (2011)
## Results: POS Tagging

<table>
<thead>
<tr>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
</table>

**Number of Languages with Best Results**

- Full model

**Average Accuracy**

(Without tag dictionary)
## Results: POS Tagging

### Number of Languages with Best Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Languages</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Gradient (DG)</td>
<td>2</td>
<td>(Portuguese, Danish)</td>
</tr>
<tr>
<td>Uniform + DG</td>
<td>2</td>
<td>(Turkish, Bulgarian)</td>
</tr>
<tr>
<td>Mixture + DG</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

### Average Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Gradient (DG)</td>
<td>40.6</td>
</tr>
<tr>
<td>Uniform + DG</td>
<td>41.0</td>
</tr>
<tr>
<td>Mixture + DG</td>
<td>43.3</td>
</tr>
</tbody>
</table>

(Without tag dictionary)

Cohen, Das and Smith (2011)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*EMNLP 2011: Cohen, Das and Smith (2011)*
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Monolingual EM (Klein and Manning, 2004)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td></td>
</tr>
<tr>
<td>PR</td>
<td></td>
</tr>
<tr>
<td>PGI</td>
<td></td>
</tr>
</tbody>
</table>

Posterior Regularization (Gillenwater et al, 2010)

Cohen, Das and Smith (2011)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ** Phylogenetic Grammar Induction (Berg-Kirkpatrick and Klein, 2010)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Uniform mixture parameters
2. No coarse-to-fine expansion (no learning)
Results: Dependency Parsing

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Languages with Best Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Learned mixture parameters
2. No coarse-to-fine expansion

Cohen, Das and Smith (2011)
## Results: Dependency Parsing

### Number of Languages with Best Results

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Uniform mixture parameters
2. Coarse-to-fine expansion → monolingual learning

Cohen, Das and Smith (2011)

EMNLP 2011
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td></td>
</tr>
<tr>
<td>PR</td>
<td></td>
</tr>
<tr>
<td>PGI</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Learned mixture parameters
2. Coarse-to-fine expansion → monolingual learning
## Results: Dependency Parsing

Cohen, Das and Smith (2011)

<table>
<thead>
<tr>
<th>Number of Languages with Best Results</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turkish, Slovene</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Accuracy</td>
<td>41.4</td>
<td>50.2*</td>
<td>53.6*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Results: Dependency Parsing

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Languages</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>with Best Results</td>
<td></td>
<td>(Turkish, Slovene)</td>
<td></td>
<td>(Bulgarian, Swedish, Dutch)</td>
<td>(Danish)</td>
<td>(Greek)</td>
<td>(Portuguese, Japanese, Spanish)</td>
</tr>
<tr>
<td>Average Accuracy</td>
<td>41.4</td>
<td>50.2*</td>
<td>53.6*</td>
<td>61.6</td>
<td>62.2</td>
<td>61.5</td>
<td>62.1</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
Analyzing $\beta$ with Principal Component Analysis

Dim 1 (34.73%)

Dim 2 (21.97%)

Japanese

Greek

Spanish

Portuguese

Dutch

Danish

Bulgarian

Swedish

Slovene
Analyzing \( \beta \) with Principal Component Analysis

Two principal components
Analyzing $\beta$ with Principal Component Analysis

Two principal components
From Words to Dependencies
Use induced tags to induce dependencies

1. In a pipeline
2. Using the posteriors over tags in a sausage lattice

(Cohen and Smith, 2007)
From Words to Dependencies

**Joint Decoding:**

$$\text{tree}^* = \arg \max_{\text{tree, POSs}} p(\text{POSs, tree}) \cdot \prod_i p(\text{POS at i | sentence})$$
From Words to Dependencies

Joint Decoding:

$$\text{tree}^* = \arg \max_{\text{tree, POSs}} p(\text{POSs, tree}) \cdot \prod_{i} p(\text{POS at i} \mid \text{sentence})$$
From Words to Dependencies

Joint Decoding:

\[
\text{DMV} = \arg\max_{\text{tree, POSs}} p(\text{POSs}, \text{tree}) \prod_i p(\text{POS at } i \mid \text{sentence})
\]

The Skibo Castle

1. DET: 0.95
   - ADJ: 0.03
   - NOUN: 0.02

2. DET: 0.0
   - ADJ: 0.3
   - NOUN: 0.7

3. DET: 0.01
   - ADJ: 0.1
   - NOUN: 0.89

Cohen, Das and Smith (2011)
From Words to Dependencies

**Joint Decoding:**

```
1  
DET: 0.95
ADJ: 0.03
NOUN: 0.02

2  
ADJ: 0.3
NOUN: 0.7

3  
ADJ: 0.1
NOUN: 0.89

4  
DET: 0.0
DET: 0.01

The Skibo Castle
```

`DMV` 

```
tree* = arg max_{tree, POSs} p(POSs, tree) \prod_i p(POS at i | sentence)
```
## Results: Words to Dependencies

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DG</strong></td>
<td>Mixture  + DG</td>
<td>DG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Number of Languages with Best Results</strong></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results: Words to Dependencies

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>DG</td>
<td>DG</td>
</tr>
<tr>
<td>Mixture + DG</td>
<td>Mixture + DG</td>
</tr>
</tbody>
</table>

### Number of Languages with Best Results

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td>1 (Greek)</td>
<td>0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>56.9</td>
<td>54.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td>5 (Portuguese, Turkish, Swedish, Slovebe, Danish)</td>
<td>4 (Bulgarian, Japanese, Spanish, Dutch)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>57.9</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
Results: Words to Dependencies

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DG</td>
<td>DG</td>
</tr>
<tr>
<td>Number of Languages with Best Results</td>
<td>1 (Greek)</td>
<td>5 (Portuguese, Turkish, Swedish, Slovak, Danish)</td>
</tr>
<tr>
<td>Average</td>
<td>56.9</td>
<td>57.9</td>
</tr>
</tbody>
</table>

Best average result with gold tags: 62.2
## Results: Words to Dependencies

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th></th>
<th>Joint</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DG</td>
<td>Mixture + DG</td>
<td>DG</td>
<td>Mixture + DG</td>
</tr>
<tr>
<td><strong>Number of Languages with Best Results</strong></td>
<td>1 (Greek)</td>
<td>0</td>
<td>5 (Portuguese, Turkish, Swedish, Slovene, Danish)</td>
<td>4 (Bulgarian, Japanese, Spanish, Dutch)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>56.9</td>
<td>54.0</td>
<td>57.9</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Best average result with gold tags: 62.2

**Interesting result:** Auto tags perform better for **Turkish and Slovene**

Cohen, Das and Smith (2011)
Conclusions
Conclusions

• Improvements for two major tasks using non-parallel multilingual guidance

• In general grammar induction results better than POS tagging
Conclusions

• Improvements for two major tasks using non-parallel multilingual guidance

• In general grammar induction results better than POS tagging

• Joint POS and dependency parsing performs surprisingly well
  • For a few languages, results are better than using gold tags
  • Joint decoding performs better than a pipeline
Questions?
### Results: POS Tagging

<table>
<thead>
<tr>
<th>Language</th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>34.7</td>
<td>38.0</td>
<td>35.8</td>
</tr>
<tr>
<td>Danish</td>
<td>48.8</td>
<td>36.2</td>
<td>39.9</td>
</tr>
<tr>
<td>Dutch</td>
<td>45.4</td>
<td>43.7</td>
<td>50.2</td>
</tr>
<tr>
<td>Greek</td>
<td>35.3</td>
<td>36.7</td>
<td>38.9</td>
</tr>
<tr>
<td>Japanese</td>
<td>52.3</td>
<td>60.4</td>
<td>61.7</td>
</tr>
<tr>
<td>Portuguese</td>
<td>53.5</td>
<td>45.7</td>
<td>51.5</td>
</tr>
<tr>
<td>Slovene</td>
<td>33.4</td>
<td>35.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>40.0</td>
<td>31.8</td>
<td>40.5</td>
</tr>
<tr>
<td>Swedish</td>
<td>34.4</td>
<td>37.7</td>
<td>39.9</td>
</tr>
<tr>
<td>Turkish</td>
<td>27.9</td>
<td>43.6</td>
<td>38.6</td>
</tr>
<tr>
<td>Average</td>
<td>40.6</td>
<td>41.0</td>
<td>43.3</td>
</tr>
</tbody>
</table>

(Without tag dictionary)

Cohen, Das and Smith (2011)
## Results: POS Tagging

<table>
<thead>
<tr>
<th>Language</th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>34.7</td>
<td>38.0</td>
<td>35.8</td>
</tr>
<tr>
<td>Danish</td>
<td>48.8</td>
<td>36.2</td>
<td>39.9</td>
</tr>
<tr>
<td>Dutch</td>
<td>45.4</td>
<td>43.7</td>
<td>50.2</td>
</tr>
<tr>
<td>Greek</td>
<td>35.3</td>
<td>36.7</td>
<td>38.9</td>
</tr>
<tr>
<td>Japanese</td>
<td>52.3</td>
<td>60.4</td>
<td>61.7</td>
</tr>
<tr>
<td>Portuguese</td>
<td>53.5</td>
<td>45.7</td>
<td>51.5</td>
</tr>
<tr>
<td>Slovene</td>
<td>33.4</td>
<td>35.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>40.0</td>
<td>31.8</td>
<td>40.5</td>
</tr>
<tr>
<td>Swedish</td>
<td>34.4</td>
<td>37.7</td>
<td>39.9</td>
</tr>
<tr>
<td>Turkish</td>
<td>27.9</td>
<td>43.6</td>
<td>38.6</td>
</tr>
<tr>
<td>Average</td>
<td>40.6</td>
<td>41.0</td>
<td>43.3</td>
</tr>
</tbody>
</table>

(without tag dictionary)

Cohen, Das and Smith (2011)
## Results: POS Tagging

<table>
<thead>
<tr>
<th>Language</th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>80.7</td>
<td>81.3</td>
<td>82.6</td>
</tr>
<tr>
<td>Danish</td>
<td>82.3</td>
<td>82.0</td>
<td>82.0</td>
</tr>
<tr>
<td>Dutch</td>
<td>79.2</td>
<td>79.3</td>
<td>80.0</td>
</tr>
<tr>
<td>Greek</td>
<td>88.0</td>
<td>80.3</td>
<td>80.3</td>
</tr>
<tr>
<td>Japanese</td>
<td>83.4</td>
<td>77.9</td>
<td>79.9</td>
</tr>
<tr>
<td>Portuguese</td>
<td>75.4</td>
<td>83.8</td>
<td>84.7</td>
</tr>
<tr>
<td>Slovene</td>
<td>75.6</td>
<td>82.8</td>
<td>82.8</td>
</tr>
<tr>
<td>Spanish</td>
<td>82.3</td>
<td>82.3</td>
<td>83.3</td>
</tr>
<tr>
<td>Swedish</td>
<td>61.5</td>
<td>69.0</td>
<td>67.0</td>
</tr>
<tr>
<td>Turkish</td>
<td>50.4</td>
<td>50.4</td>
<td>50.4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>75.9</strong></td>
<td><strong>76.9</strong></td>
<td><strong>77.3</strong></td>
</tr>
</tbody>
</table>

(with tag dictionary)

Cohen, Das and Smith (2011)
## Results: POS Tagging

<table>
<thead>
<tr>
<th>Language</th>
<th>Direct Gradient (DG)</th>
<th>Uniform + DG</th>
<th>Mixture + DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>80.7</td>
<td>81.3</td>
<td><strong>82.6</strong></td>
</tr>
<tr>
<td>Danish</td>
<td><strong>82.3</strong></td>
<td>82.0</td>
<td>82.0</td>
</tr>
<tr>
<td>Dutch</td>
<td>79.2</td>
<td>79.3</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>Greek</td>
<td><strong>88.0</strong></td>
<td>80.3</td>
<td>80.3</td>
</tr>
<tr>
<td>Japanese</td>
<td>83.4</td>
<td>77.9</td>
<td>79.9</td>
</tr>
<tr>
<td>Portuguese</td>
<td>75.4</td>
<td>83.8</td>
<td><strong>84.7</strong></td>
</tr>
<tr>
<td>Slovene</td>
<td>75.6</td>
<td><strong>82.8</strong></td>
<td>82.8</td>
</tr>
<tr>
<td>Spanish</td>
<td>82.3</td>
<td>82.3</td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>Swedish</td>
<td>61.5</td>
<td><strong>69.0</strong></td>
<td>67.0</td>
</tr>
<tr>
<td>Turkish</td>
<td><strong>50.4</strong></td>
<td><strong>50.4</strong></td>
<td>50.0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>75.9</strong></td>
<td><strong>76.9</strong></td>
<td><strong>77.3</strong></td>
</tr>
</tbody>
</table>

(with tag dictionary)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Language</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>54.3</td>
<td>54.0</td>
<td>-</td>
<td>75.6</td>
<td>75.5</td>
<td>74.7</td>
<td>72.8</td>
</tr>
<tr>
<td>Danish</td>
<td>41.4</td>
<td>44.0</td>
<td>41.6</td>
<td>59.2</td>
<td>59.9</td>
<td>51.3</td>
<td>55.2</td>
</tr>
<tr>
<td>Dutch</td>
<td>38.6</td>
<td>37.9</td>
<td>45.1</td>
<td>50.7</td>
<td>51.1</td>
<td>45.9</td>
<td>46.0</td>
</tr>
<tr>
<td>Greek</td>
<td>41.0</td>
<td>-</td>
<td>-</td>
<td>57.0</td>
<td>59.5</td>
<td>73.0</td>
<td>72.3</td>
</tr>
<tr>
<td>Japanese</td>
<td>43.0</td>
<td>60.2</td>
<td>-</td>
<td>56.3</td>
<td>58.3</td>
<td>59.8</td>
<td>63.9</td>
</tr>
<tr>
<td>Portuguese</td>
<td>42.5</td>
<td>47.8</td>
<td>63.1</td>
<td>78.6</td>
<td>76.8</td>
<td>78.7</td>
<td>79.8</td>
</tr>
<tr>
<td>Slovene</td>
<td>37.0</td>
<td>50.3</td>
<td>49.6</td>
<td>46.1</td>
<td>46.0</td>
<td>41.3</td>
<td>41.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>38.1</td>
<td>62.4</td>
<td>63.8</td>
<td>73.2</td>
<td>75.9</td>
<td>75.5</td>
<td>76.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>42.3</td>
<td>42.2</td>
<td>58.3</td>
<td>74.0</td>
<td>73.2</td>
<td>70.5</td>
<td>68.7</td>
</tr>
<tr>
<td>Turkish</td>
<td>36.3</td>
<td>53.4</td>
<td>-</td>
<td>45.0</td>
<td>45.3</td>
<td>43.9</td>
<td>44.1</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>41.4</td>
<td>-</td>
<td>-</td>
<td>61.6</td>
<td>62.2</td>
<td>61.5</td>
<td>62.1</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
## Results: Dependency Parsing

<table>
<thead>
<tr>
<th>Language</th>
<th>EM</th>
<th>PR</th>
<th>PGI</th>
<th>Uniform</th>
<th>Mixture</th>
<th>Uniform + EM</th>
<th>Mixture + EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>54.3</td>
<td>54.0</td>
<td>-</td>
<td><strong>75.6</strong></td>
<td>75.5</td>
<td>74.7</td>
<td>72.8</td>
</tr>
<tr>
<td>Danish</td>
<td>41.4</td>
<td>44.0</td>
<td>41.6</td>
<td>59.2</td>
<td><strong>59.9</strong></td>
<td>51.3</td>
<td>55.2</td>
</tr>
<tr>
<td>Dutch</td>
<td>38.6</td>
<td>37.9</td>
<td>45.1</td>
<td>50.7</td>
<td><strong>51.1</strong></td>
<td>45.9</td>
<td>46.0</td>
</tr>
<tr>
<td>Greek</td>
<td>41.0</td>
<td>-</td>
<td>-</td>
<td>57.0</td>
<td>59.5</td>
<td><strong>73.0</strong></td>
<td>72.3</td>
</tr>
<tr>
<td>Japanese</td>
<td>43.0</td>
<td>60.2</td>
<td>-</td>
<td>56.3</td>
<td>58.3</td>
<td>59.8</td>
<td><strong>63.9</strong></td>
</tr>
<tr>
<td>Portuguese</td>
<td>42.5</td>
<td>47.8</td>
<td>63.1</td>
<td>78.6</td>
<td>76.8</td>
<td>78.7</td>
<td><strong>79.8</strong></td>
</tr>
<tr>
<td>Slovene</td>
<td>37.0</td>
<td>50.3</td>
<td>49.6</td>
<td>46.1</td>
<td>46.0</td>
<td>41.3</td>
<td>41.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>38.1</td>
<td>62.4</td>
<td>63.8</td>
<td>73.2</td>
<td>75.9</td>
<td>75.5</td>
<td><strong>76.7</strong></td>
</tr>
<tr>
<td>Swedish</td>
<td>42.3</td>
<td>42.2</td>
<td>58.3</td>
<td><strong>74.0</strong></td>
<td>73.2</td>
<td>70.5</td>
<td>68.7</td>
</tr>
<tr>
<td>Turkish</td>
<td>36.3</td>
<td>53.4</td>
<td>-</td>
<td>45.0</td>
<td>45.3</td>
<td>43.9</td>
<td>44.1</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>41.4</td>
<td>-</td>
<td>-</td>
<td><strong>61.6</strong></td>
<td><strong>62.2</strong></td>
<td><strong>61.5</strong></td>
<td><strong>62.1</strong></td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)
## Results: Words to Dependencies

<table>
<thead>
<tr>
<th>Language</th>
<th>Joint DG</th>
<th>Joint Mixture + DG</th>
<th>Pipeline DG</th>
<th>Pipeline Mixture + DG</th>
<th>Gold Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>62.4</td>
<td>67.0</td>
<td>57.7</td>
<td>62.9</td>
<td>75.6</td>
</tr>
<tr>
<td>Danish</td>
<td>50.4</td>
<td>50.1</td>
<td>48.9</td>
<td>48.3</td>
<td>59.9</td>
</tr>
<tr>
<td>Dutch</td>
<td>48.3</td>
<td>52.2</td>
<td>49.9</td>
<td>51.2</td>
<td>50.7</td>
</tr>
<tr>
<td>Greek</td>
<td>63.5</td>
<td>52.2</td>
<td>68.2</td>
<td>50.0</td>
<td>73.0</td>
</tr>
<tr>
<td>Japanese</td>
<td>61.4</td>
<td>69.5</td>
<td>64.2</td>
<td>68.6</td>
<td>63.9</td>
</tr>
<tr>
<td>Portuguese</td>
<td>68.4</td>
<td>62.2</td>
<td>60.0</td>
<td>59.8</td>
<td>79.8</td>
</tr>
<tr>
<td>Slovene</td>
<td>47.2</td>
<td>36.8</td>
<td>45.8</td>
<td>36.4</td>
<td>46.1</td>
</tr>
<tr>
<td>Spanish</td>
<td>67.7</td>
<td>69.3</td>
<td>65.8</td>
<td>68.1</td>
<td>76.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>58.2</td>
<td>49.1</td>
<td>57.9</td>
<td>47.6</td>
<td>74.0</td>
</tr>
<tr>
<td>Turkish</td>
<td>52.4</td>
<td>47.4</td>
<td>50.8</td>
<td>47.1</td>
<td>45.3</td>
</tr>
<tr>
<td>Average</td>
<td>57.9</td>
<td>55.0</td>
<td>56.9</td>
<td>54.0</td>
<td>64.5</td>
</tr>
</tbody>
</table>

Cohen, Das and Smith (2011)