Optimizing Spectral Learning for Parsing

Shashi Narayan, Shay Cohen
School of Informatics, University of Edinburgh

ACL, August 2016
Probabilistic CFGs with Latent States (Matsuzaki et al., 2005; Prescher 2005)

Latent states play the role of nonterminal subcategorization, e.g., \( \text{NP} \rightarrow \{ \text{NP}^1, \text{NP}^2, \ldots, \text{NP}^{24} \} \)

- analogous to syntactic heads as in lexicalization (Charniak 1997)?

They are not part of the observed data in the treebank
Estimating PCFGs with Latent States (L-PCFGs)

**EM Algorithm** *(Matsuzaki et al., 2005; Petrov et al., 2006)*

↓ **Problems with local maxima**; it fails to provide certain type of theoretical guarantees as it doesn’t find global maximum of the log-likelihood
Estimating PCFGs with Latent States (L-PCFGs)

**EM Algorithm** *(Matsuzaki et al., 2005; Petrov et al., 2006)*

\[\downarrow\] **Problems with local maxima**; it fails to provide certain type of theoretical guarantees as it doesn’t find global maximum of the log-likelihood

**Spectral Algorithm** *(Cohen et al., 2012, 2014)*

\[\uparrow\] Statistically consistent algorithms that make use of spectral decomposition

\[\uparrow\] Much faster training than the EM algorithm
Estimating PCFGs with Latent States (L-PCFGs)

**EM Algorithm** (Matsuzaki et al., 2005; Petrov et al., 2006)

- Problems with local maxima; it fails to provide certain type of theoretical guarantees as it doesn’t find global maximum of the log-likelihood

**Spectral Algorithm** (Cohen et al., 2012, 2014)

- Statistically consistent algorithms that make use of spectral decomposition
- Much faster training than the EM algorithm
- Lagged behind in their empirical results
Overview


Conventional approach: Number of latent states for each nonterminal in an L-PCFG can be decided in isolation
Overview


Conventional approach: Number of latent states for each nonterminal in an L-PCFG can be decided in isolation

Contributions:

A. Parsing results significantly improve if the number of latent states for each nonterminal is globally optimized

- Petrov et al. (2006) demonstrated that coarse-to-fine techniques that carefully select the number of latent states improve accuracy.
Overview


Conventional approach: Number of latent states for each nonterminal in an L-PCFG can be decided in isolation

Contributions:

B. Optimized spectral method beats coarse-to-fine expectation-maximization techniques on 6 (Basque, Hebrew, Hungarian, Korean, Polish and Swedish) out of 8 SPMRL datasets
Intuition behind the Spectral Algorithm

Inside and outside trees

At node VP:

Outside tree $o =$

Inside tree $t =$

Conditionally independent given the label and the hidden state

$$p(o, t|VP, h) = p(o|VP, h) \times p(t|VP, h)$$
Recent Advances in Spectral Estimation

Singular value decomposition (SVD) of cross-covariance matrix for each nonterminal
Recent Advances in Spectral Estimation

Method of moments (Cohen et al., 2012, 2014)

- Averaging with SVD parameters ⇒ Dense estimates
Recent Advances in Spectral Estimation

Method of moments (Cohen et al., 2012, 2014)

- Averaging with SVD parameters ⇒ Dense estimates

Clustering variants (Narayan and Cohen 2015)

Sparse estimates
Standard Spectral Estimation and Number of Latent States

A natural way to choose the number of latent states based on the number of non-zero singular values

Number of latent states for each nonterminal in an L-PCFG can be decided in isolation

Conventional approach fails to take into account interactions between different nonterminals
Optimizing Latent States for Various Nonterminals

Input:

- An input treebank divided into training and development set
- A basic spectral estimation algorithm $S$ mapping each nonterminal to a fixed number of latent states
  - $f_{\text{def}} : \{ S \rightarrow 24, \ NNP \rightarrow 24, \ VP \rightarrow 24, \ DT \rightarrow 24, \ldots \}$

Output:

- $f_{\text{opt}} : \{ S \rightarrow 40, \ NNP \rightarrow 81, \ VP \rightarrow 35, \ DT \rightarrow 4, \ldots \}$
Algorithm in a nutshell

- Iterate through the nonterminals, changing the number of latent states,
- estimate the grammar on the training set and
- optimize the accuracy on the dev set

A beam search algorithm for the traversal of multidimensional vectors of latent states: Optimizing their global interaction
Optimizing Latent States for Various Nonterminals

$time: 0$

\[
\begin{array}{c}
\text{DT} \\
\text{S} \\
\text{NP}
\end{array}
\]

\[f_{\text{def}} : 24 \quad 24 \quad 24 \quad 24 \quad 24 \quad 24 \quad , \quad F_{1_{\text{def}}}
\]
Optimizing Latent States for Various Nonterminals

\[ f_{\text{def}} : \begin{array}{cccc} 4 & 37 & 24 & 24 \\ \end{array}, F_{1_{\text{def}}} \]

\[ f_{m_1} : \begin{array}{cccc} 4 & 37 & m_1 & 24 \end{array}, F_{1_{m_1}} \]

\[ f_{m_2} : \begin{array}{cccc} 4 & 37 & m_2 & 24 \end{array}, F_{1_{m_2}} \]

\[ f_{m_N} : \begin{array}{cccc} 4 & 37 & m_N & 24 \end{array}, F_{1_{m_N}} \]
Optimizing Latent States for Various Nonterminals

Clustering variant of spectral estimation leads to compact models and is relatively fast.
Experiments

The SPMRL Dataset

8 morphologically rich languages: Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish

Treebanks of varying sizes from 5,000 sentences (Hebrew and Swedish) to 40,472 sentences (German)
Results on the Swedish dataset

Results on the dev set

![Bar chart showing F-Measures for different methods on the dev set.]

- **berkeley cluster moments**
  - Petrov et al.'06: 75.50
  - Björkelund et al.'13

- **cluster**
  - Narayan and Cohen'15: 71.40

- **moments**
  - Cohen et al.'13: 73.40
Results on the Swedish dataset

Results on the dev set

<table>
<thead>
<tr>
<th>Method</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>berkeley</td>
<td>75.50</td>
</tr>
<tr>
<td>cluster</td>
<td>71.40</td>
</tr>
<tr>
<td>moments</td>
<td>75.20</td>
</tr>
<tr>
<td>Petrov et al.'06</td>
<td></td>
</tr>
<tr>
<td>Narayan and Cohen'15</td>
<td></td>
</tr>
<tr>
<td>Cohen et al.'13</td>
<td></td>
</tr>
</tbody>
</table>

Björkelund et al.'13
Results on the Swedish dataset

Results on the dev set

<table>
<thead>
<tr>
<th>Method</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petrov et al.'06</td>
<td>75.50</td>
</tr>
<tr>
<td>Narayan and Cohen'15</td>
<td>71.40</td>
</tr>
<tr>
<td>Cohen et al.'13</td>
<td>75.20</td>
</tr>
<tr>
<td>Björkelund et al.'13</td>
<td>73.40</td>
</tr>
<tr>
<td>Berkeley Cluster Moments</td>
<td>75.50</td>
</tr>
</tbody>
</table>
Results on the Swedish dataset

Final results on the test set

- berkeley: 80.60
- cluster: 76.40, 79.40
- moments: 78.40, 80.90

Authors:
- Petrov et al.'06
- Narayan and Cohen'15
- Cohen et al.'13
- Björkelund et al.'13
Final Results on the SPMRL Dataset

![Bar chart showing F-Measures for different languages and methods.]

- Basque
- French
- German
- Hebrew
- Hungarian
- Korean
- Polish
- Swedish

Berkeley results are taken from Björkelund et al, 2013.
Conclusion

Spectral parsing results significantly improve if the number of latent states for each nonterminal is globally optimized

Optimized spectral algorithm beats coarse-to-fine EM algorithm for 6 (Basque, Hebrew, Hungarian, Korean, Polish and Swedish) out of 8 SPMRL datasets

The Rainbow parser and multilingual models:
http://cohort.inf.ed.ac.uk/lpcfg/

Acknowledgments: Thanks to David McClosky, Eugene Charniak, DK Choe, Geoff Gordon, Djamé Seddah, Thomas Müller, Anders Björkelund and anonymous reviewers
Inside Features used

Consider the VP node in the following tree:

The inside features consist of:
- The pairs \((\text{VP}, \text{V})\) and \((\text{VP}, \text{NP})\)
- The rule \(\text{VP} \rightarrow \text{V NP}\)
- The tree fragment \((\text{VP} (\text{V saw}) \text{NP})\)
- The tree fragment \((\text{VP} \text{V} (\text{NP} \text{D N}))\)
- The pair of head part-of-speech tag with \(\text{VP}: (\text{VP}, \text{V})\)
Consider the D node in the following tree:

The outside features consist of:

- The pairs \((D, NP)\) and \((D, NP, VP)\)
- The pair of head part-of-speech tag with \(D: (D, N)\)
- The tree fragments \(NP, VP\), and
Variants of Spectral Estimation

- **SVD variants:** singular value decomposition of empirical count matrices (cross-covariance matrices) to estimate grammar parameters (Cohen et. al. 2012, 2014)

- **Convex EM variant:** “anchor method” that identifies features that uniquely identify latent states (Cohen and Collins, 2014)

- **Clustering variant:** a simplified version of the SVD variant that clusters low-dimensional representations to latent states (Narayan and Cohen, 2015)

**Intuitive-to-understand and very (computationally) efficient**
Optimizing Latent States for Various Nonterminals

▶ **Initialization:** \( (n_0, f_{\text{def}}, F_{\text{def}}) \rightarrow Q \)
  - \( n_0 \): First nonterminal
  - \( f_{\text{def}} \): \( \{S \rightarrow 24, \text{NNP} \rightarrow 24, \text{VP} \rightarrow 24, \text{DT} \rightarrow 24, \ldots\} \)
  - \( F_{\text{def}} \) is the \( F_1 \) score on the development set

▶ **Iteration:** \( (n_i, f_i, F_i) \leftarrow Q \)
  - For each number of latent state \( l \in \{1, \ldots, m\} \),
  - \( f'_i \): \( f'_i(n_i) = l \) and for others \( n \), \( f'_i(n) = f_i(n) \),
  - Estimate a new \( F'_i \) score on the development set, and
  - Push \( (n_{i+1}, f'_i, F'_i) \)

▶ **Termination:** \( (n_{\text{fin}+1}, f_{\text{opt}}, F_{\text{fin}}) \leftarrow Q \)
  - \( f_{\text{opt}} \): \( \{S \rightarrow 40, \text{NNP} \rightarrow 81, \text{VP} \rightarrow 35, \text{DT} \rightarrow 4, \ldots\} \)

We need a training algorithm which is relatively fast and leads to compact models.
## Final Results on the SPMRL Dataset

<table>
<thead>
<tr>
<th>Language</th>
<th>Berkeley</th>
<th>Spectral Cluster</th>
<th>Spectral SVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basque</td>
<td>74.7</td>
<td><strong>81.4</strong></td>
<td>80.5</td>
</tr>
<tr>
<td>French</td>
<td>80.4</td>
<td>75.6</td>
<td><strong>79.1</strong></td>
</tr>
<tr>
<td>German</td>
<td>78.3</td>
<td>76.0</td>
<td><strong>78.2</strong></td>
</tr>
<tr>
<td>Hebrew</td>
<td>87.0</td>
<td>87.2</td>
<td><strong>89.0</strong></td>
</tr>
<tr>
<td>Hungarian</td>
<td>85.2</td>
<td>88.4</td>
<td><strong>89.2</strong></td>
</tr>
<tr>
<td>Korean</td>
<td>78.6</td>
<td>78.4</td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>Polish</td>
<td>86.8</td>
<td><strong>91.2</strong></td>
<td>91.8</td>
</tr>
<tr>
<td>Swedish</td>
<td>80.6</td>
<td>79.4</td>
<td><strong>80.9</strong></td>
</tr>
</tbody>
</table>
Spectral Algorithm Vs Treebank Size

We break the common belief that more data is needed with spectral algorithm

<table>
<thead>
<tr>
<th>lang.</th>
<th>Training data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sent.</td>
<td>tokens</td>
</tr>
<tr>
<td>Basque</td>
<td>7,577</td>
<td>96,565</td>
</tr>
<tr>
<td>French</td>
<td>14,759</td>
<td>443,113</td>
</tr>
<tr>
<td>German</td>
<td>40,472</td>
<td>719,532</td>
</tr>
<tr>
<td>Hebrew</td>
<td>5,000</td>
<td>128,065</td>
</tr>
<tr>
<td>Hungarian</td>
<td>8,146</td>
<td>170,221</td>
</tr>
<tr>
<td>Korean</td>
<td>23,010</td>
<td>301,800</td>
</tr>
<tr>
<td>Polish</td>
<td>6,578</td>
<td>66,814</td>
</tr>
<tr>
<td>Swedish</td>
<td>5,000</td>
<td>76,332</td>
</tr>
<tr>
<td>lang.</td>
<td>$\sum_{nt} Is_{nt}$ Before</td>
<td>$\sum_{nt} Is_{nt}$ After</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Basque</td>
<td>402</td>
<td>646</td>
</tr>
<tr>
<td>French</td>
<td>1984</td>
<td>1994</td>
</tr>
<tr>
<td>German</td>
<td>2288</td>
<td>2213</td>
</tr>
<tr>
<td>Hebrew</td>
<td>603</td>
<td>986</td>
</tr>
<tr>
<td>Hungarian</td>
<td>643</td>
<td>676</td>
</tr>
<tr>
<td>Korean</td>
<td>1295</td>
<td>1200</td>
</tr>
<tr>
<td>Polish</td>
<td>384</td>
<td>491</td>
</tr>
<tr>
<td>Swedish</td>
<td>276</td>
<td>629</td>
</tr>
</tbody>
</table>
Multilingual Models for the Rainbow Parser

The Rainbow Parser (or RParser) is a phrase-structure syntactic parser developed at the University of Edinburgh by the informal research group Cohort. At its core, the use of a latent-variable PCFG model. Its training procedure is based on spectral methods of learning. The parser is not publicly available yet. However, if you are interested in using it for your research, contact Shay Cohen (scohen AT inf.ed.ac.uk) or Shashi Narayan (snaraya2 AT inf.ed.ac.uk).

Click for the following paper.

```latex
@inproceedings{narayan-16b,
  title={Optimizing Spectral Learning for Parsing},
  author={Shashi Narayan and Shay B. Cohen},
  booktitle={Proceedings of {ACL}},
  year={2016}
}
```

Below we include the table of results on the test sets from the SPMRL shared task to parse morphologically rich languages. For a legend, see the paper (Tables 2 and 3).

<table>
<thead>
<tr>
<th>Language</th>
<th>CL van.</th>
<th>CL opt.</th>
<th>SP van.</th>
<th>SP opt.</th>
<th>Berkeley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basque</td>
<td>79.6</td>
<td>81.4</td>
<td>79.9</td>
<td>80.5</td>
<td>74.7</td>
</tr>
<tr>
<td>French</td>
<td>74.3</td>
<td>75.6</td>
<td>78.7</td>
<td>79.1</td>
<td>80.4</td>
</tr>
<tr>
<td>German (NEGRA)</td>
<td>76.4</td>
<td>78.0</td>
<td>78.4</td>
<td>79.4</td>
<td>80.1</td>
</tr>
<tr>
<td>German (TiGeR)</td>
<td>74.1</td>
<td>76.0</td>
<td>78.0</td>
<td>78.2</td>
<td>78.3</td>
</tr>
<tr>
<td>Hebrew</td>
<td>86.3</td>
<td>87.2</td>
<td>87.8</td>
<td>89.0</td>
<td>87.0</td>
</tr>
<tr>
<td>Hungarian</td>
<td>86.5</td>
<td>88.4</td>
<td>89.1</td>
<td>89.2</td>
<td>85.2</td>
</tr>
<tr>
<td>Korean</td>
<td>76.5</td>
<td>78.4</td>
<td>80.3</td>
<td>80.0</td>
<td>78.6</td>
</tr>
<tr>
<td>Polish</td>
<td>90.5</td>
<td>91.2</td>
<td>91.8</td>
<td>91.8</td>
<td>86.8</td>
</tr>
<tr>
<td>Swedish</td>
<td>76.4</td>
<td>79.4</td>
<td>78.4</td>
<td>80.9</td>
<td>80.6</td>
</tr>
</tbody>
</table>