Optimizing Spectral Learning for Parsing

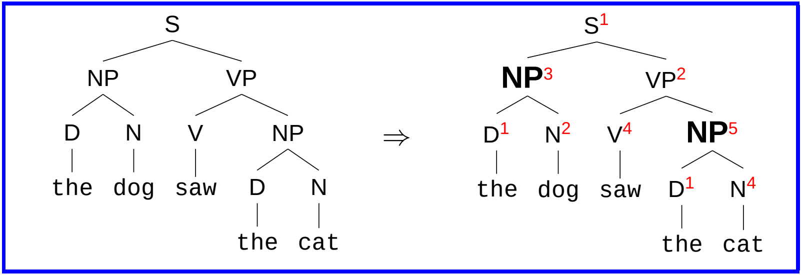
Shashi Narayan, Shay Cohen

School of Informatics, University of Edinburgh

ACL, August 2016

THE UNIVERSITY of EDINBURGH

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



Latent states play the role of nonterminal subcategorization, e.g., 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)

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

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)

- $\Downarrow$ **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
- $\Downarrow$ **Lagged behind in their empirical results**

Overview

Builds on the work on the spectral algorithm for Latent-state PCFGs (L-PCFGs) for parsing (Cohen et al., 2012, 2014, Cohen and Collins, 2014, Narayan and Cohen 2015)

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

Overview

Builds on the work on the spectral algorithm for Latent-state PCFGs (L-PCFGs) for parsing (Cohen et al., 2012, 2014, Cohen and Collins, 2014, Narayan and Cohen 2015)

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

Builds on the work on the spectral algorithm for Latent-state PCFGs (L-PCFGs) for parsing (Cohen et al., 2012, 2014, Cohen and Collins, 2014, Narayan and Cohen 2015)

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

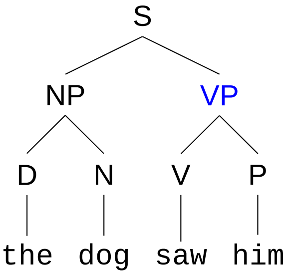
Contributions:

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

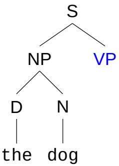
Intuition behind the Spectral Algorithm

Inside and outside trees

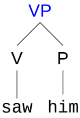
At node VP:



Outside tree $o =$



Inside tree $t =$



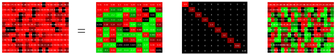
Conditionally independent given the label and the hidden state

$$p(o, t | \text{VP}, h) = p(o | \text{VP}, h) \times p(t | \text{VP}, h)$$

Recent Advances in Spectral Estimation

Singular value decomposition (SVD) of cross-covariance matrix for each nonterminal

Recent Advances in Spectral Estimation

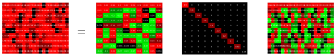


SVD Step

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

- Averaging with SVD parameters ⇒ Dense estimates

Recent Advances in Spectral Estimation

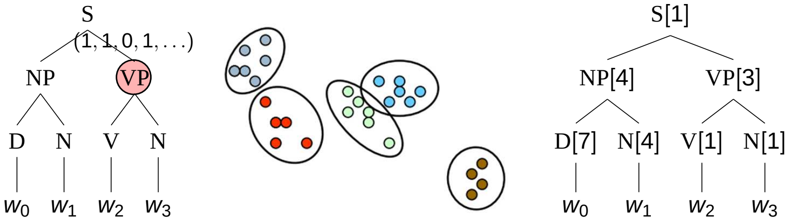


SVD Step

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
- $f_{def} : \{\text{S} \rightarrow 24, \text{NNP} \rightarrow 24, \text{VP} \rightarrow 24, \text{DT} \rightarrow 24, \ldots\}$

Output:

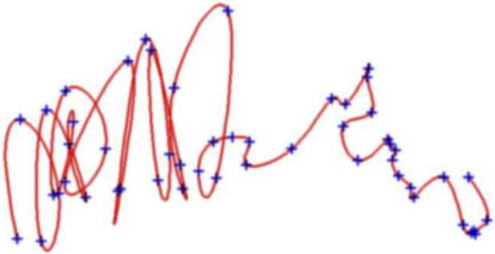
- $f_{opt} : \{\text{S} \rightarrow 40, \text{NNP} \rightarrow 81, \text{VP} \rightarrow 35, \text{DT} \rightarrow 4, \ldots\}$

Optimizing Latent States for Various Nonterminals

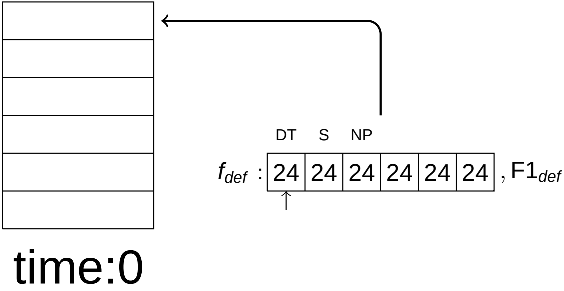
Algorithm in a nutshell

- 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

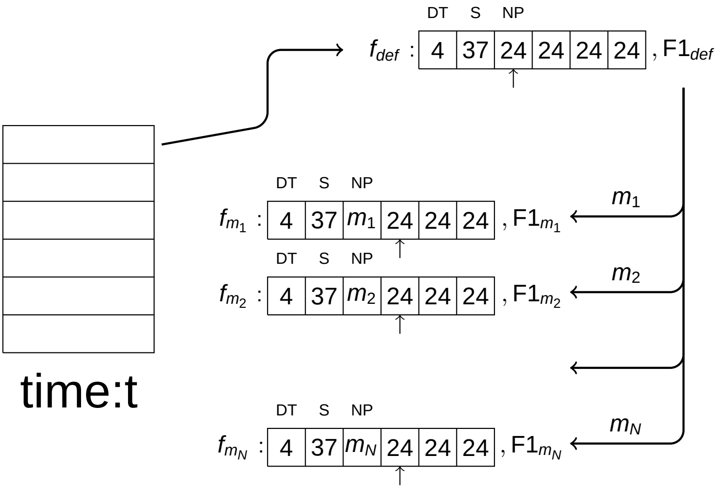
A **beam search algorithm** for the traversal of multidimensional vectors of latent states: **Optimizing their global interaction**



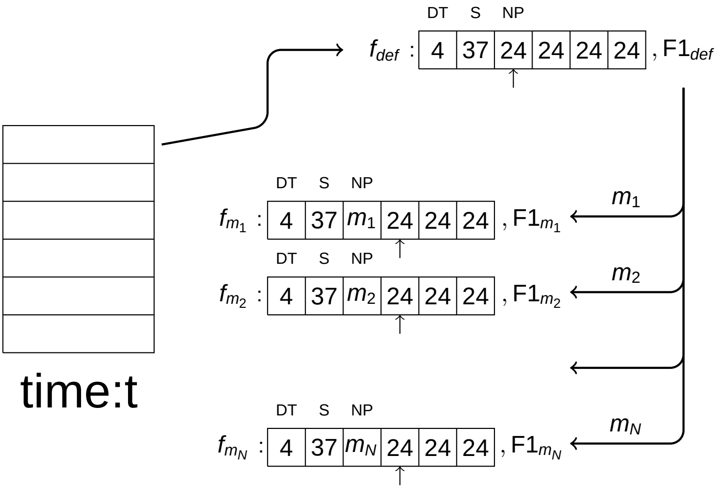
Optimizing Latent States for Various Nonterminals



Optimizing Latent States for Various Nonterminals



Optimizing Latent States for Various Nonterminals



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

Experiments

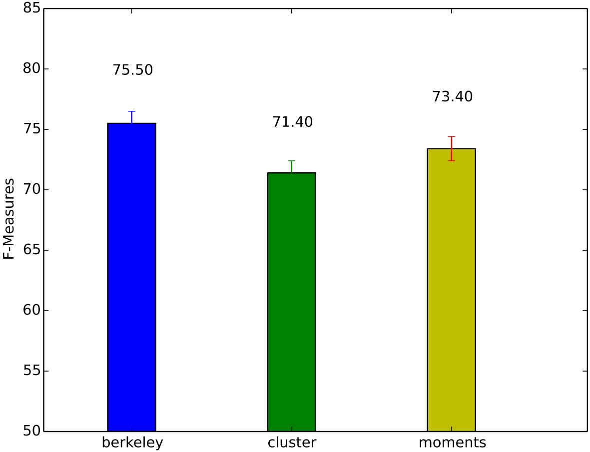
The SPMRL Dataset

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

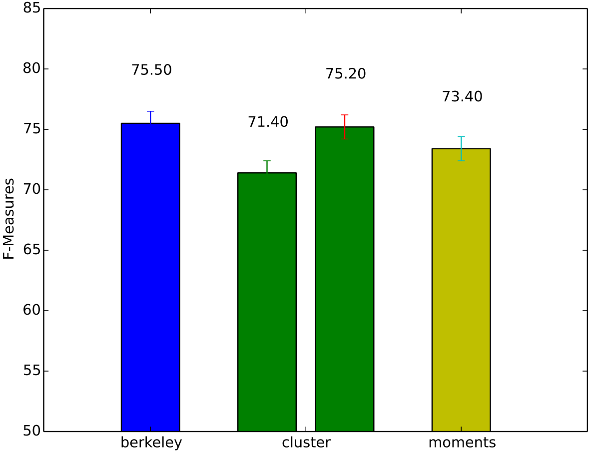


Petrov et al.'06 Narayan and Cohen'15 Cohen et al.'13

Björkelund et al.'13

Results on the Swedish dataset

Results on the dev set

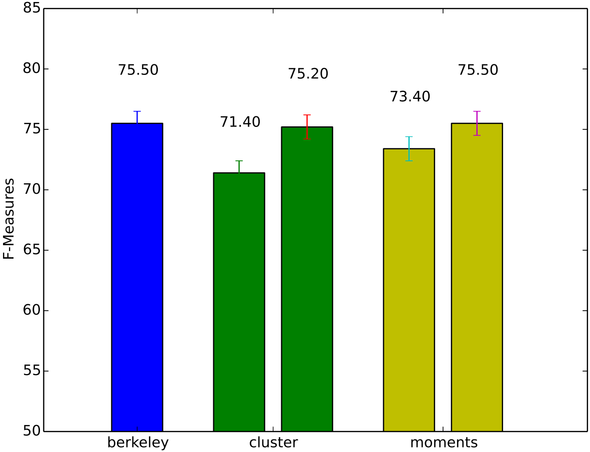


Petrov et al.'06 Narayan and Cohen'15 Cohen et al.'13

Björkelund et al.'13

Results on the Swedish dataset

Results on the dev set

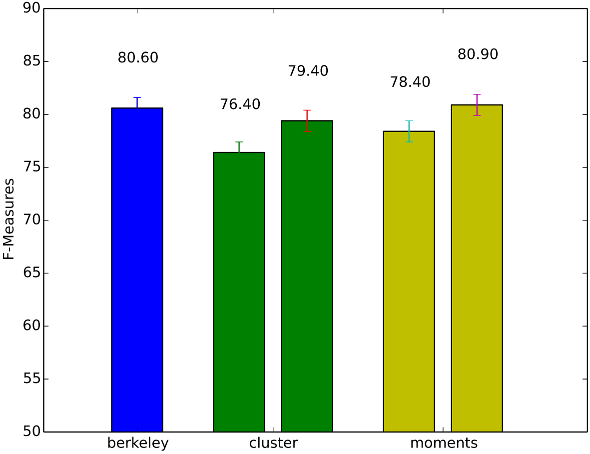


Petrov et al.'06 Narayan and Cohen'15 Cohen et al.'13

Björkelund et al.'13

Results on the Swedish dataset

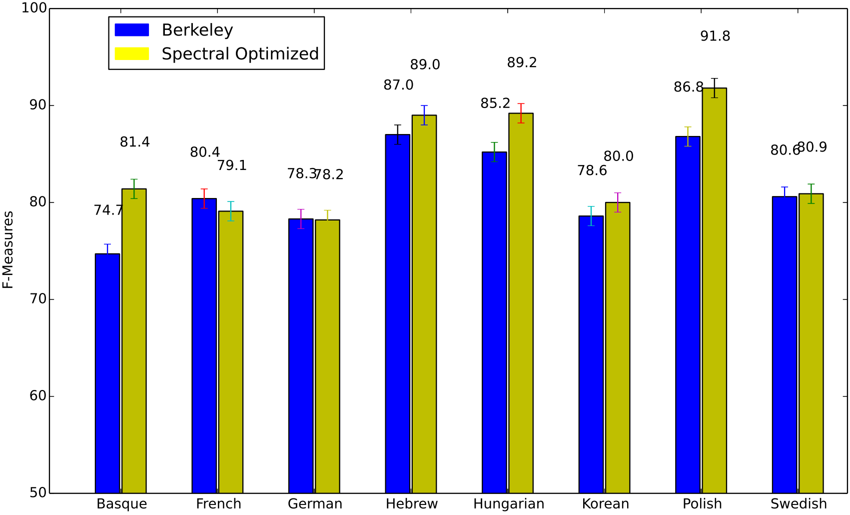
Final results on the test set



Petrov et al.'06 Narayan and Cohen'15 Cohen et al.'13

Björkelund et al.'13

Final Results on the SPMRL Dataset



- 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

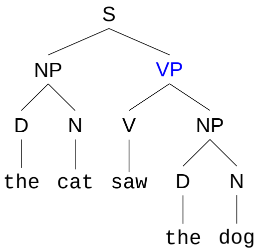
Optimized spectral algorithm beats coarse-to-fine EM algorithm for 6 (**Basque**, **Hebrew**, **Hungarian**, **Korean**, **Polish** and **Swedish**) out of 8 **SPMRL** datasets

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:

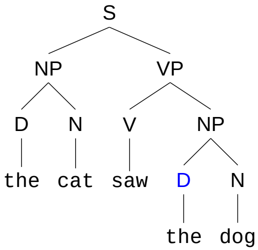


The inside features consist of:

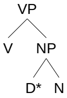
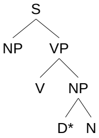
- The pairs `(VP, V)` and `(VP, NP)`
- The rule VP $\rightarrow$ `V NP`
- The tree fragment `(VP (V saw) NP)`
- The tree fragment `(VP V (NP D N))`
- The pair of head part-of-speech tag with `VP: (VP, V)`

Outside Features used

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  ,  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:** $(n_0, f_{def}, F_{def}) \rightarrow Q$
  - n_0 : First nonterminal
  - f_{def} : $\{$S $\rightarrow$ 24, NNP $\rightarrow$ 24, VP $\rightarrow$ 24, DT $\rightarrow$ 24, $\ldots\}$
  - F_{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 $l \in \{1, ..., m\}$,
  - 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:** $(n_{fin+1}, f_{opt}, F_{fin}) \leftarrow Q$
  - f_{opt} : $\{$S$\rightarrow$ 40, NNP $\rightarrow$ 81, VP $\rightarrow$ 35, DT $\rightarrow$ 4, $\ldots\}$

We need a training algorithm which is relatively fast and leads to compact models

Final Results on the SPMRL Dataset

lang.	Berkeley	Spectral	
		Cluster	SVD
Basque	74.7	**<u>81.4</u>**	80.5
French	<u>80.4</u>	75.6	**79.1**
German	<u>78.3</u>	76.0	**78.2**
Hebrew	87.0	87.2	**<u>89.0</u>**
Hungarian	85.2	88.4	**<u>89.2</u>**
Korean	78.6	78.4	**<u>80.0</u>**
Polish	86.8	**<u>91.2</u>**	91.8
Swedish	80.6	79.4	**<u>80.9</u>**

Spectral Algorithm Vs Treebank Size

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

lang.	Training data	
	Sent.	tokens
Basque	7,577	96,565
French	14,759	443,113
German	40,472	719,532
Hebrew	5,000	128,065
Hungarian	8,146	170,221
Korean	23,010	301,800
Polish	6,578	66,814
Swedish	5,000	76,332

Effect of Optimization on the Model Size

lang.	$\sum_{nt} ls_{nt}$		#nt
	Before	After	
Basque	402	646	200
French	1984	1994	222
German	**2288**	**2213**	762
Hebrew	603	986	375
Hungarian	643	676	112
Korean	**1295**	**1200**	352
Polish	384	491	198
Swedish	276	629	148

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.

```
@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).

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