Optimizing Spectral Learning for Parsing

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
We describe a search algorithm for optimizing the number of latent states when estimating latent-variable PCFGs with spectral methods. Our results show that contrary to the common belief that the number of latent states for each nonterminal in an L-PCFG can be decided in isolation with spectral methods, parsing results significantly improve if the number of latent states for each nonterminal is globally optimized, while taking into account interactions between the different nonterminals. In addition, we contribute an empirical analysis of spectral algorithms on eight morphologically rich languages: Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish. Our results show that our estimation consistently performs better or close to coarse-to-fine expectation-maximization techniques for these languages.

1 Introduction

Latent-variable probabilistic context-free grammars (L-PCFGs) have been used in the natural language processing community (NLP) for syntactic parsing for over a decade. They were introduced in the NLP community by Matsuzaki et al. (2005) and Prescher (2005), with Matsuzaki et al. using the expectation-maximization (EM) algorithm to estimate them. Their performance on syntactic parsing of English at that stage lagged behind state-of-the-art parsers.

Petrov et al. (2006) showed that one of the reasons that the EM algorithm does not estimate state-of-the-art parsing models for English is that the EM algorithm does not control well for the model size used in the parser – the number of latent states associated with the various nonterminals in the grammar. As such, they introduced a coarse-to-fine technique to estimate the grammar. It splits and merges nonterminals (with latent state information) with the aim to optimize the likelihood of the training data. Together with other types of fine tuning of the parsing model, this led to state-of-the-art results for English parsing.

In more recent work, Cohen et al. (2012) described a different family of estimation algorithms for L-PCFGs. This so-called “spectral” family of learning algorithms is compelling because it offers a rigorous theoretical analysis of statistical convergence, and sidesteps local maxima issues that arise with the EM algorithm.

While spectral algorithms for L-PCFGs are compelling from a theoretical perspective, they have been lagging behind in their empirical results on the problem of parsing. In this paper we show that one of the main reasons for that is that spectral algorithms require a more careful tuning procedure for the number of latent states than that which has been advocated for until now. In a sense, the relationship between our work and the work of Cohen et al. (2013) is analogous to the relationship between the work by Petrov et al. (2006) and the work by Matsuzaki et al. (2005): we suggest a technique for optimizing the number of latent states for spectral algorithms, and test it on eight languages.

Our results show that when the number of latent states is optimized using our technique, the parsing models the spectral algorithms yield perform significantly better than the vanilla-estimated models, and for most of the languages – better than the Berkeley parser of Petrov et al. (2006).

As such, the contributions of this parser are two-fold:

- We describe a search algorithm for optimiz-
ing the number of latent states for spectral learning.

- We describe an analysis of spectral algorithms on eight languages (until now the results of L-PCFG estimation with spectral algorithms for parsing were known only for English). Our parsing algorithm is rather language-generic, and does not require significant linguistically-oriented adjustments.

In addition, we dispel the common wisdom that more data is needed with spectral algorithms. Our models yield high performance on treebanks of varying sizes from 5,000 sentences (Hebrew and Swedish) to 40,472 sentences (German).

The rest of the paper is organized as follows. In §2 we describe notation and background. §3 further investigates the need for an optimization of the number of latent states in spectral learning and describes our optimization algorithm, a search algorithm akin to beam search. In §4 we describe our experiments with natural language parsing for Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish. We conclude in §5.

2 Background and Notation

We denote by $[n]$ the set of integers $\{1, \ldots, n\}$. An L-PCFG is a 5-tuple $(\mathcal{N}, \mathcal{I}, \mathcal{P}, f, n)$ where:

- $\mathcal{N}$ is the set of nonterminal symbols in the grammar. $\mathcal{I} \subset \mathcal{N}$ is a finite set of internals. $\mathcal{P} \subset \mathcal{N}$ is a finite set of preterminals. We assume that $\mathcal{N} = \mathcal{I} \cup \mathcal{P}$, and $\mathcal{I} \cap \mathcal{P} = \emptyset$. Hence we have partitioned the set of nonterminals into two subsets.

- $f: \mathcal{N} \rightarrow \mathbb{N}$ is a function that maps each nonterminal $a$ to the number of latent states it uses. The set $[m_a]$ includes the possible hidden states for nonterminal $a$.

- $[n]$ is the set of possible words.

- For all $a \in \mathcal{I}$, $b \in \mathcal{N}$, $c \in \mathcal{N}$, $h_1 \in [m_a]$, $h_2 \in [m_b]$, $h_3 \in [m_c]$, we have a binary context-free rule $a(h_1) \rightarrow b(h_2) c(h_3)$.

- For all $a \in \mathcal{P}$, $h \in [m_a]$, $x \in [n]$, we have a lexical context-free rule $a(h) \rightarrow x$.

The estimation of an L-PCFG requires an assignment of probabilities (or weights) to each of the rules $a(h_1) \rightarrow b(h_2) c(h_3)$ and $a(h) \rightarrow x$, and also an assignment of starting probabilities for each $a(h)$, where $a \in \mathcal{I}$ and $h \in [m_a]$. Estimation is usually assumed to be done from a set of parse trees (a treebank), where the latent states are not included in the data – only the “skeletal” trees which consist of nonterminals in $\mathcal{N}$.

L-PCFGs, in their symbolic form, are related to regular tree grammars, an old grammar formalism, but they were introduced as statistical models for parsing with latent heads more recently by Matsuzaki et al. (2005) and Prescher (2005). Earlier work about L-PCFGs by Matsuzaki et al. (2005) used the expectation-maximization (EM) algorithm to estimate the grammar probabilities. Indeed, given that the latent states are not observed, EM is a good fit for L-PCFG estimation, since it aims to do learning from incomplete data. This work has been further extended by Petrov et al. (2006) to use EM in a coarse-to-fine fashion: merging and splitting nonterminals using the latent states to optimize the number of latent states for each nonterminal.

Cohen et al. (2012) presented a so-called spectral algorithm to estimate L-PCFGs. This algorithm uses linear-algebraic procedures such as singular value decomposition (SVD) during learning. The spectral algorithm of Cohen et al. builds on an estimation algorithm for HMMs by Hsu et al. (2009). Cohen et al. (2013) experimented with this spectral algorithm for parsing English. A different variant of a spectral learning algorithm for L-PCFGs was developed by Cohen and Collins (2014). It breaks the problem of L-PCFG estimation into multiple convex optimization problems which are solved using EM.

The family of L-PCFG spectral learning algorithms was further extended by Narayan and Cohen (2015). They presented a simplified version of the algorithm of Cohen et al. (2012) that estimates sparse grammars and assigns probabilities (instead of weights) to the rules in the grammar, and as such does not suffer from the problem of negative probabilities that arise with the original spectral algorithm (see discussion in Cohen et al., 2013). In this paper, we use the algorithms by Narayan and Cohen (2015) and Cohen

\footnote{A related algorithm for weighted tree automata (WTA) was developed by Bailly et al. (2010). However, the conversion from L-PCFGs to WTA is not straightforward, and information is lost in this conversion. See also (Rabusseau et al., 2016).}
et al. (2012), and we compare them against state-of-the-art L-PCFG parsers such as the Berkeley parser (Petrov et al., 2006). We also compare our algorithms to other state-of-the-art parsers where elaborate linguistically-motivated feature specifications (Hall et al., 2014), annotations (Crabbé, 2015) and formalism conversions (Fernández-González and Martins, 2015) are used.

3 Optimizing Spectral Estimation

In this section, we describe our optimization algorithm and its motivation.

3.1 Spectral Learning of L-PCFGs and Model Size

The family of spectral algorithms for latent-variable PCFGs rely on feature functions that are defined for inside and outside trees. Given a tree, the inside tree for a node contains the entire subtree below that node; the outside tree contains everything in the tree excluding the inside tree. Figure 1 shows an example of inside and outside trees for the nonterminal VP in the parse tree of the sentence “the mouse chased the cat”.

![Figure 1: The inside tree (left) and outside tree (right) for the nonterminal VP in the parse tree (S (NP (D the) (N mouse)) (VP (V chased) (NP (D the) (N cat)))) for the sentence “the mouse chased the cat.”](image)

With L-PCFGs, the model dictates that an inside tree and an outside tree that are connected at a node are statistically conditionally independent of each other given the node label and the latent state that is associated with it. As such, one can identify the distribution over the latent states for a given nonterminal \(a\) by using the cross-covariance matrix of the inside and the outside trees, \(\Omega^a\). For more information on the definition of this cross-covariance matrix, see Cohen et al. (2012) and Narayan and Cohen (2015).

The L-PCFG spectral algorithms use singular value decomposition (SVD) on \(\Omega^a\) to reduce the dimensionality of the feature functions. If \(\Omega^a\) is computed from the true L-PCFG distribution then the rank of \(\Omega^a\) (the number of non-zero singular values) gives the number of latent states according to the model.

In the case of estimating \(\Omega^a\) from data generated from an L-PCFG, the number of latent states for each nonterminal can be exposed by capping it when the singular values of \(\Omega^a\) are smaller than some threshold value. This means that spectral algorithms give a natural way for the selection of the number of latent states for each nonterminal. This is where our algorithm comes into play: it optimizes the number of latent search for each nonterminal by applying a search algorithm akin to beam search.

3.2 Optimizing the Number of Latent States

As mentioned in the previous section, the number of non-zero singular values of \(\Omega^a\) gives a criterion to determine the number of latent states \(m_a\) for a given nonterminal \(a\). In practice, we cap \(m_a\) not to include small singular values which are close to 0, because of estimation errors of \(\Omega^a\).

This procedure does not take into account the interactions that exist between choices of latent state numbers for the various nonterminals. In principle, given the independence assumptions that L-PCFGs make, choosing the nonterminals based only on the singular values is “statistically correct.” However, because in practice the modeling assumptions that we make (that natural language parse trees are drawn from an L-PCFG) do not hold, we can improve further the accuracy of the model by taking into account the nonterminal interaction. Another source of difficulty in choosing the number of latent states based the singular values of \(\Omega^a\) is sampling error: in practice, we are using data to estimate \(\Omega^a\), and as such, even if the model is correct, the rank of the estimated matrix does not have to correspond to the rank of \(\Omega^a\) according to the true distribution. As a matter of fact, in addition to neglecting small singular values, the spectral methods of Cohen et al. (2013) and Narayan and Cohen (2015) also cap the number of latent states for each nonterminal to an up-
**Algorithm:**

(Step 0: Initialization)

- Set $Q$, a queue of size $k$, to be empty.
- Estimate an L-PCFG $G_S : (\mathcal{N}, \mathcal{I}, \mathcal{P}, f_S, n)$ using $\mathcal{S}$.
- Initialize $f = f_S$, a function that maps each nonterminal $a \in \mathcal{N}$ to the number of latent states.
- Let $L$ be a list of nonterminals $(a_1, \ldots, a_M)$ such that $a_i \in \mathcal{N}$ for which to optimize the number of latent states.
- Let $s$ be the $F_1$ score for the above L-PCFG $G_S$ on the development set.
- Put in $Q$ the element $(s, 1, f, \text{coarse})$.
- The queue is ordered by $s$, the first element of tuples, in the queue.

(Step 1: Search, repeat until termination happens)

- Dequeue the queue into $(s, j, f, t)$ where $j$ is the index in the input nonterminal list $L$. 
- If $j = (M + 1)$, return $f$.
- If $t$ is coarse then for each $m_0 \in \{1, 5, 10, \ldots, m\}$:
  - Let $f_0$ be such that $\forall a \neq a_j f_0(a) = f(a)$ and $f_0(a_j) = m_0$.
  - Train an L-PCFG $G_0$ using $\mathcal{S}$ but with $f_0$.
  - Let $s_0$ be the $F_1$ score for $G_0$ on the development set.
  - Enqueue into $Q$: $(s_0, j, f_0, \text{refine})$.
- If $t$ is refine then for each $m_0 \in \{f(a) + \ell \mid \ell \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}\}$:
  - Let $f_0$ be such that $\forall a \neq a_j f_0(a) = f(a)$ and $f_0(a_j) = m_0$.
  - Train an L-PCFG $G_0$ using $\mathcal{S}$ but with $f_0$.
  - Let $s_0$ be the $F_1$ score for $G_0$ on the development set.
  - Enqueue into $Q$: $(s_0, j + 1, f_0, \text{coarse})$.

Figure 2: A search algorithm for finding the optimal number of latent states.

Petrov et al. (2006) improves over the estimation described in Matsuzaki et al. (2005) by taking into account the interactions between the nonterminals and their latent state numbers in the training data. They use the EM algorithm to split and merge nonterminals using the latent states, and optimize the number of latent states for each nonterminal such that it maximizes the likelihood of a training treebank. Their refined grammar successfully splits nonterminals to various degrees to capture their complexity. We take the analogous step with spectral methods. We propose an algorithm where we first compute $\Omega^a$ on the training data and then we optimize the number of latent states for each nonterminal by optimizing the PARSEVAL metric (Black et al., 1991) on a development set.

Our optimization algorithm appears in Figure 2. The input to the algorithm is training and development data in the form of parse trees, a basic spectral estimation algorithm $\mathcal{S}$ in its default setting, an upper bound $m$ on the number of latent states that can be used for the different nonterminals and a beam size $k$ which gives a maximal queue size for the beam. The algorithm aims to learn a function $f$ that maps each nonterminal $a$ to the number of latent states. It initializes $f$ by estimating a default grammar $G_S : (\mathcal{N}, \mathcal{I}, \mathcal{P}, f_S, n)$ using $\mathcal{S}$ and setting $f = f_S$. It then iterates over $a \in \mathcal{N}$, improving $f$ such that it optimizes the PARSEVAL metric on the development set.

The state of the algorithm includes a queue that consists of tuples of the form $(s, j, f, t)$ where $f$ is an assignment of latent state numbers to each nonterminal in the grammar, $j$ is the index of a nonterminal to be explored in the input nonterminal list $L$, $s$ is the $F_1$ score on the development set for a grammar that is estimated with $f$ and $t$ is a tag that can either be coarse or refine.

The algorithm orders these tuples by $s$ in the queue, and iteratively dequeues elements from the queue. Then, depending on the label $t$, it either makes a refined search for the number of latent states for $a_j$, or a more coarse search. As such, the algorithm can be seen as a variant of a beam search algorithm.

The search algorithm can be used with any training algorithm for L-PCFGs, including the algorithms of Cohen et al. (2013) and Narayan and Cohen (2015). These methods, in their default setting, use a function $f_S$ which maps each nonterminal $a$ to a fixed number of latent states $m_a$ it uses. In this case, $\mathcal{S}$ takes as input training data, in the form of a treebank, decomposes into inside and outside trees at each node in each tree in the training set; and reduces the dimensionality of the inside and outside feature functions by running.
SVD on the cross-covariance matrix $\Omega_a$ of the inside and the outside trees, for each nonterminal $a$. Cohen et al. (2013) estimate the parameters of the L-PCFG up to a linear transformation using $f(a)$ non-zero singular values of $\Omega_a$, whereas Narayan and Cohen (2015) use the feature representations induced from the SVD step to cluster instances of nonterminal $a$ in the training data into $f(a)$ clusters; these clusters are then treated as latent states that are “observed.” Finally, Narayan and Cohen follow up with a simple frequency count maximum likelihood estimate to estimate the parameters in the L-PCFG with these latent states.

An important point to make is that the learning algorithms of Narayan and Cohen (2015) and Cohen et al. (2013) are relatively fast, in comparison to the EM algorithm. They require only one iteration over the data. In addition, the SVD step of $S$ for these learning algorithms is computed just once for a large $m$. The SVD of a lower rank can then be easily computed from that SVD.

### 4 Experiments

In this section, we describe our setup for parsing experiments on a range of languages.

#### 4.1 Experimental Setup

**Datasets** We experiment with nine treebanks consisting of eight different morphologically rich languages: Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish. Table 1 shows the statistics of 9 different treebanks with their splits into training, development and test sets. Eight out of the nine datasets (Basque, French, German-T, Hebrew, Hungarian, Korean, Polish and Swedish) are taken from the workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL; Seddah et al., 2013). The German corpus in the SPMRL workshop is taken from the TiGer corpus (German-T, Brants et al., 2004). We also experiment with another German corpus, the NEGRA corpus (German-N, Skut et al., 1997), in a standard evaluation split. Words in the SPMRL datasets are annotated with their morphological signatures, whereas the NEGRA corpus does not contain any morphological information.

**Data preprocessing and treatment of rare words** We convert all trees in the treebanks to a binary form, train and run the parser in that form, and then transform back the trees when doing evaluation using the PARSEVAL metric. In addition, we collapse unary rules into unary chains, so that our trees are fully binarized. The column “#nts” in Table 1 shows the number of nonterminals after binarization in the various treebanks. Before binarization, we also drop all functional information from the nonterminals. We use fine tags for all languages except Korean. This is in line with Björkelund et al. (2013). For Korean, there are 2,825 binarized nonterminals making it impractical to use our optimization algorithm, so we use the coarse tags. Björkelund et al. (2013) have shown that the morphological signatures for rare words are useful to improve the performance of the Berkeley parser.

<table>
<thead>
<tr>
<th>lang</th>
<th>Basque</th>
<th>French</th>
<th>German-N</th>
<th>German-T</th>
<th>Hebrew</th>
<th>Hungarian</th>
<th>Korean</th>
<th>Polish</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>sent.</td>
<td>7,577</td>
<td>14,759</td>
<td>18,602</td>
<td>40,472</td>
<td>5,000</td>
<td>8,146</td>
<td>23,010</td>
<td>6,578</td>
<td>5,000</td>
</tr>
<tr>
<td>tokens</td>
<td>96,565</td>
<td>443,113</td>
<td>328,531</td>
<td>719,532</td>
<td>128,065</td>
<td>170,221</td>
<td>301,800</td>
<td>66,814</td>
<td>76,332</td>
</tr>
<tr>
<td>lex. size</td>
<td>25,136</td>
<td>27,470</td>
<td>48,509</td>
<td>77,219</td>
<td>18,602</td>
<td>11,305</td>
<td>76,332</td>
<td>17,002</td>
<td>18,602</td>
</tr>
<tr>
<td>#nts</td>
<td>112</td>
<td>222</td>
<td>208</td>
<td>762</td>
<td>375</td>
<td>112</td>
<td>198</td>
<td>148</td>
<td>148</td>
</tr>
<tr>
<td>sent.</td>
<td>948</td>
<td>1,235</td>
<td>5,000</td>
<td>5,000</td>
<td>5,000</td>
<td>3,010</td>
<td>2,066</td>
<td>821</td>
<td>494</td>
</tr>
<tr>
<td>tokens</td>
<td>13,893</td>
<td>38,820</td>
<td>76,704</td>
<td>11,305</td>
<td>1,051</td>
<td>25,729</td>
<td>8,391</td>
<td>9,339</td>
<td>9,339</td>
</tr>
<tr>
<td>sent.</td>
<td>946</td>
<td>2,541</td>
<td>5,000</td>
<td>716</td>
<td>7,400</td>
<td>19,913</td>
<td>28,783</td>
<td>8,336</td>
<td>666</td>
</tr>
<tr>
<td>tokens</td>
<td>11,477</td>
<td>75,216</td>
<td>28,783</td>
<td>17,002</td>
<td>11,477</td>
<td>76,332</td>
<td>40,775</td>
<td>30,010</td>
<td>28,783</td>
</tr>
</tbody>
</table>

Table 1: Statistics about the different datasets used in our experiments for the training (“train”), development (“dev”) and test (“test”) sets. “sent.” denotes the number of sentences in the dataset, “tokens” denotes the total number of words in the dataset, “lex. size” denotes the vocabulary size in the training set and “#nts” denotes the number of nonterminals in the training set after binarization.
In our preliminary experiments with naive spectral estimation, we preprocess rare words in the training set in two ways: (i) we replace them with their corresponding POS tags, and (ii) we replace them with their corresponding POS+morphological signatures. We follow Björkelund et al. (2013) and consider a word to be rare if it occurs less than 20 times in the training data. We experimented both with a version of the parser that does not ignore and does ignore letter cases, and discovered that the parser behaves better when case is not ignored.

**Spectral algorithms: subroutine choices** The latent state optimization algorithm will work with either the clustering estimation algorithm of Narayan and Cohen (2015) or the spectral algorithm of Cohen et al. (2013). In our setup, we first run the latent state optimization algorithm with the clustering algorithm. We then run the spectral algorithm once with the optimized $f$ from the clustering algorithm. We do that because the clustering algorithm is significantly faster to iteratively parse the development set, because it leads to sparse estimates.

Our optimization algorithm is sensitive to the initialization of the number of latent states assigned to each nonterminal as it sequentially goes through the list of nonterminals and chooses latent state numbers for each nonterminal, keeping latent state numbers for other nonterminals fixed. In our setup, we start our search algorithm with the best model from the clustering algorithm, controlling for all hyperparameters; we tune $f$, the function which maps each nonterminal to a fixed number of latent states $m$, by running the vanilla version with different values of $m$ for different languages. Based on our preliminary experiments, we set $m$ to 4 for Basque, Hebrew, Polish and Swedish; 8 for German-N; 16 for German-T, Hungarian and Korean; and 24 for French.

We use the same features for the spectral methods as in Narayan and Cohen (2015) for German-N. For the SPMRL datasets we do not use the head features. These require linguistic understanding of the datasets (because they require head rules for propagating leaf nodes in the tree), and we discovered that simple heuristics for constructing these rules did not yield an increase in performance.

We use the `kmeans` function in Matlab to do the clustering for the spectral algorithm of Narayan and Cohen (2015). We experimented with several versions of $k$-means, and discovered that the version that works best in a set of preliminary experiments is hard $k$-means.5

**Decoding and evaluation** For efficiency, we use a base PCFG without latent states to prune marginals which receive a value less than 0.00005 in the dynamic programming chart. This is just a bare-bones PCFG that is estimated using maximum likelihood estimation (with frequency count). The parser takes part-of-speech tagged sentences as input. We tag the German-N data using the Turbo Tagger (Martins et al., 2010). For the languages in the SPMRL data we use the MarMot tagger of Müller et al. (2013) to jointly predict the POS and morphological tags.6 The parser itself can assign different part-of-speech tags to words to avoid parse failure. This is also particularly important for constituency parsing with morphologically rich languages. It helps mitigate the problem of the taggers to assign correct tags when long-distance dependencies are present.

For all results, we report the $F_1$ measure of the PARSEVAL metric (Black et al., 1991). We use the EVALB program7 with the parameter file Collins.prm (Collins, 1999) for the German-N data and the SPMRL parameter file, spmrl.prm, for the SPMRL data (Seddah et al., 2013).

In this setup, the latent state optimization algorithm terminates in few hours for all datasets except French and German-T. The German-T data has 762 nonterminals to tune over a large development set consisting of 5,000 sentences, whereas, the French data has a high average sentence length of 31.43 in the development set.8

Following Narayan and Cohen (2015), we further improve our results by using multiple spectral models where noise is added to the underlying features in the training set before the estimation of

---

5To be more precise, we use the Matlab function `kmeans` while passing it the parameter `start='sample'` to randomly sample the initial centroid positions. In our experiments, we found that default initialization of centroids differs in Matlab14 (random) and in Matlab15 (kmeans++). Our estimation performs better with random initialization.

6See Björkelund et al. (2013) for the performance of the MarMot tagger on the SPMRL datasets.


8To speed up tuning on the French data, we drop sentences with length >46 from the development set, dropping its size from 12,35 to 1,006.
Table 2: Results on the development datasets. “Bk” makes use of the Berkeley parser with its coarse-to-fine mechanism to optimize the number of latent states (Petrov et al., 2006). For Bk, “van” uses the vanilla treatment of rare words using signatures defined by Petrov et al. (2006), whereas “rep.” uses the morphological signatures instead. “Cl” uses the algorithm of Narayan and Cohen (2015) and “Sp” uses the algorithm of Cohen et al. (2013). In Cl, “van (pos)” and “van (rep)” are vanilla estimations (i.e., each nonterminal is mapped to fixed number of latent states) replacing rare words by POS or POS+morphological signatures, respectively. The best of these two models is used with our optimization algorithm in “opt”. For Sp, “van” uses the best setting for unknown words as Cl. Best result in each column from the first seven rows is in bold. In addition, our best performing models from rows 3-7 are marked with “∗”.

<table>
<thead>
<tr>
<th>lang</th>
<th>Basque</th>
<th>French</th>
<th>German-N</th>
<th>German-T</th>
<th>Hebrew</th>
<th>Hungarian</th>
<th>Korean</th>
<th>Polish</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van</td>
<td>69.2</td>
<td>79.9</td>
<td>81.7</td>
<td>87.8</td>
<td>83.9</td>
<td>71.0</td>
<td>84.1</td>
<td>74.5</td>
<td></td>
</tr>
<tr>
<td>rep</td>
<td>84.3</td>
<td>79.7</td>
<td>82.7</td>
<td>89.6</td>
<td>89.1</td>
<td>82.8</td>
<td>87.1</td>
<td>75.5</td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van (pos)</td>
<td>69.8</td>
<td>73.9</td>
<td>75.7</td>
<td>80.0</td>
<td>81.3</td>
<td>68.7</td>
<td>90.3</td>
<td>70.9</td>
<td></td>
</tr>
<tr>
<td>van (rep)</td>
<td>78.6</td>
<td>73.7</td>
<td>78.8</td>
<td>88.1</td>
<td>84.7</td>
<td>76.5</td>
<td>90.4</td>
<td>71.4</td>
<td></td>
</tr>
<tr>
<td>opt</td>
<td>81.2*</td>
<td>76.7</td>
<td>77.8</td>
<td>90.1</td>
<td>87.2</td>
<td>79.2</td>
<td>92.0*</td>
<td>75.2</td>
<td></td>
</tr>
<tr>
<td>Sp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van</td>
<td>78.1</td>
<td>78.0</td>
<td>77.6</td>
<td>82.0</td>
<td>89.2</td>
<td>87.7</td>
<td>80.6</td>
<td>91.7</td>
<td>73.4</td>
</tr>
<tr>
<td>opt</td>
<td>79.0</td>
<td>78.1*</td>
<td>79.0*</td>
<td>82.9*</td>
<td>90.3*</td>
<td>87.8*</td>
<td>80.9*</td>
<td>91.7</td>
<td>75.5</td>
</tr>
<tr>
<td>Bk multiple</td>
<td>87.4</td>
<td>82.5</td>
<td>85.0</td>
<td>90.5</td>
<td>91.1</td>
<td>84.6</td>
<td>88.4</td>
<td>79.5</td>
<td></td>
</tr>
<tr>
<td>Cl multiple</td>
<td>83.4</td>
<td>79.9</td>
<td>85.1</td>
<td>90.6</td>
<td>89.0</td>
<td>80.8</td>
<td>92.5</td>
<td>78.3</td>
<td></td>
</tr>
<tr>
<td>Hall et al.’14</td>
<td>83.7</td>
<td>79.4</td>
<td>83.3</td>
<td>88.1</td>
<td>87.4</td>
<td>81.9</td>
<td>91.1</td>
<td>76.0</td>
<td></td>
</tr>
<tr>
<td>Crabbé ’15</td>
<td>84.0</td>
<td>80.9</td>
<td>84.1</td>
<td>90.7</td>
<td>88.3</td>
<td>83.1</td>
<td>92.8</td>
<td>77.9</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Results on the test datasets. “Bk” denotes the best Berkeley parser result reported by the shared task organizers (Seddah et al., 2013). For the German-N data, Bk results are taken from Petrov (2010). “Cl van” shows the performance of the best vanilla models from Table 2 on the test set. “Cl opt” and “Sp opt” give the result of our algorithm on the test set. We also include results from Hall et al. (2014) and Crabbé (2015).

<table>
<thead>
<tr>
<th>lang</th>
<th>Basque</th>
<th>French</th>
<th>German-N</th>
<th>German-T</th>
<th>Hebrew</th>
<th>Hungarian</th>
<th>Korean</th>
<th>Polish</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van</td>
<td>74.7</td>
<td>80.4</td>
<td>86.1</td>
<td>78.3</td>
<td>87.0</td>
<td>85.2</td>
<td>78.6</td>
<td>86.8</td>
<td>80.6</td>
</tr>
<tr>
<td>rep</td>
<td>84.4</td>
<td>75.6</td>
<td>78.0</td>
<td>76.0</td>
<td>87.2</td>
<td>88.4</td>
<td>78.4</td>
<td>91.2</td>
<td>79.4</td>
</tr>
<tr>
<td>Cl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van (pos)</td>
<td>79.9</td>
<td>78.7</td>
<td>81.4</td>
<td>74.1</td>
<td>86.3</td>
<td>86.5</td>
<td>76.5</td>
<td>90.5</td>
<td>76.4</td>
</tr>
<tr>
<td>van (rep)</td>
<td>80.5</td>
<td>79.1*</td>
<td>79.4*</td>
<td>87.8</td>
<td>89.1</td>
<td>80.3</td>
<td>91.8</td>
<td>78.4</td>
<td></td>
</tr>
<tr>
<td>opt</td>
<td>84.1*</td>
<td>84.5</td>
<td>83.8</td>
<td>89.5</td>
<td>91.9</td>
<td>84.3</td>
<td>87.8</td>
<td>84.9</td>
<td></td>
</tr>
<tr>
<td>Bk multiple</td>
<td>87.9</td>
<td>82.9</td>
<td>84.5</td>
<td>81.3</td>
<td>89.5</td>
<td>91.9</td>
<td>84.3</td>
<td>87.8</td>
<td>84.9</td>
</tr>
<tr>
<td>Cl multiple</td>
<td>83.4</td>
<td>80.4</td>
<td>82.7</td>
<td>80.4</td>
<td>89.2</td>
<td>89.9</td>
<td>80.3</td>
<td>92.4</td>
<td>82.8</td>
</tr>
<tr>
<td>Hall et al.’14</td>
<td>83.4</td>
<td>79.7</td>
<td>78.4</td>
<td>87.2</td>
<td>88.3</td>
<td>80.2</td>
<td>90.7</td>
<td>82.0</td>
<td></td>
</tr>
<tr>
<td>F&amp;M ’15</td>
<td>85.9</td>
<td>78.8</td>
<td>78.7</td>
<td>89.0</td>
<td>88.2</td>
<td>79.3</td>
<td>91.2</td>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>Crabbé ’15</td>
<td>84.9</td>
<td>80.8</td>
<td>79.3</td>
<td>89.7</td>
<td>90.1</td>
<td>82.7</td>
<td>92.7</td>
<td>83.2</td>
<td></td>
</tr>
</tbody>
</table>

Using the optimized f, we estimate 80 models for each of noise induction mechanisms in Narayan and Cohen: Dropout, Gaussian (additive) and Gaussian (multiplicative). To decode with multiple noisy models, we train the MaxEnt reranker of Charniak and Johnson (2005). Hierarchical decoding with “maximal tree coverage” over MaxEnt models, further improves our accuracy. See Narayan and Cohen (2015) for more details on the estimation of a diverse set of models, and on decoding with them.

4.2 Results

Table 2 and Table 3 give the results for the various languages. Our main focus is on comparing the coarse-to-fine Berkeley parser (Petrov et al., 2006) to our method. However, for the sake of completeness, we also present results for other parsers, such as parsers of Hall et al. (2014), Fernández-González and Martins (2015) and Crabbé (2015). In line with Björkelund et al. (2013), our preliminary experiments with the treatment of rare words suggest that morphological features are useful for all SPMRL languages except French.

9We only use the algorithm of Narayan and Cohen (2015) for the noisy model estimation. They have shown that decoding with noisy models performs better with their sparse estimates than the dense estimates of Cohen et al. (2013).

10Implementation: https://github.com/BLLIP/bllip-parser. More specifically, we used the programs extract-spfeatures, cvlm-lbfgs and best-indices.extract-spfeatures uses head features, we bypass this for the SPMRL datasets by creating a dummy heads.cc file. cvlm-lbfgs was used with the default hyperparameters from the Makefile.
Table 4: A comparison of the number of latent states for the different nonterminals before and after running our latent state number optimization algorithm. The index \( i \) ranges over preterminals and interminals, with \( x_i \) denoting the number of latent states for nonterminal \( i \) with the vanilla version of the estimation algorithm and \( y_i \) denoting the number of latent states for nonterminal \( i \) after running the optimization algorithm. The divergence figure ("div.") is a calculation of \( \sum \| x_i - y_i \| \).

<table>
<thead>
<tr>
<th>language</th>
<th>( \sum x_i )</th>
<th>( \sum y_i )</th>
<th>div.</th>
<th>#nts</th>
<th>( \sum x_i )</th>
<th>( \sum y_i )</th>
<th>div.</th>
<th>#nts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basque</td>
<td>311</td>
<td>419</td>
<td>196</td>
<td>169</td>
<td>91</td>
<td>227</td>
<td>152</td>
<td>31</td>
</tr>
<tr>
<td>French</td>
<td>839</td>
<td>715</td>
<td>476</td>
<td>108</td>
<td>1145</td>
<td>1279</td>
<td>906</td>
<td>114</td>
</tr>
<tr>
<td>German-N</td>
<td>425</td>
<td>567</td>
<td>416</td>
<td>109</td>
<td>323</td>
<td>578</td>
<td>361</td>
<td>99</td>
</tr>
<tr>
<td>German-T</td>
<td>1251</td>
<td>890</td>
<td>795</td>
<td>378</td>
<td>1037</td>
<td>1323</td>
<td>738</td>
<td>384</td>
</tr>
<tr>
<td>Hebrew</td>
<td>434</td>
<td>442</td>
<td>182</td>
<td>279</td>
<td>169</td>
<td>544</td>
<td>393</td>
<td>96</td>
</tr>
<tr>
<td>Hungarian</td>
<td>457</td>
<td>415</td>
<td>282</td>
<td>87</td>
<td>186</td>
<td>261</td>
<td>129</td>
<td>25</td>
</tr>
<tr>
<td>Korean</td>
<td>1077</td>
<td>980</td>
<td>547</td>
<td>331</td>
<td>218</td>
<td>220</td>
<td>150</td>
<td>21</td>
</tr>
<tr>
<td>Polish</td>
<td>252</td>
<td>311</td>
<td>197</td>
<td>135</td>
<td>132</td>
<td>180</td>
<td>86</td>
<td>63</td>
</tr>
<tr>
<td>Swedish</td>
<td>191</td>
<td>284</td>
<td>127</td>
<td>106</td>
<td>85</td>
<td>345</td>
<td>266</td>
<td>42</td>
</tr>
</tbody>
</table>

Specifically, for Basque, Hungarian and Korean, improvements are significantly large.

Our results show that the optimization of the number of latent states with the clustering and spectral algorithms indeed improves these algorithms performance, and these increases generalize to the test sets as well. This was a point of concern, since the optimization algorithm goes through many points in the hypothesis space of parsing models, and identifies one that behaves optimally on the development set – and as such it could overfit to the development set. However, this did not happen, and in some cases, the increase in accuracy of the test set after running our optimization algorithm is actually larger than the one for the development set.

While the vanilla estimation algorithms (without latent state optimization) lag behind the Berkeley parser for many of the languages, once the number of latent states is optimized, our parsing models do better for Basque, Hebrew, Hungarian, Korean, Polish and Swedish. For German-T we perform close to the Berkeley parser (78.2 vs. 78.3). It is also interesting to compare the clustering algorithm of Narayan and Cohen (2015) to the spectral algorithm of Cohen et al. (2013). In the vanilla version, the spectral algorithm does better in most cases. However, these differences are narrowed, and in some cases, overcome, when the number of latent states is optimized. Decoding with multiple models further improves our accuracy. Our “Cl multiple” results lag behind “Bk multiple.” We believe this is the result of the need of head features for the MaxEnt models.\(^\text{12}\)

Our results show that spectral learning is a viable alternative to the use of expectation-maximization coarse-to-fine techniques. As we discuss later, further improvements have been introduced to state-of-the-art parsers that are orthogonal to the use of a specific estimation algorithm. Some of them can be applied to our setup.

4.3 Further Analysis

In addition to the basic set of parsing results, we also wanted to inspect the size of the parsing models when using the optimization algorithm in comparison to the vanilla models. Table 4 gives this analysis. In this table, we see that in most cases, on average, the optimization algorithm chooses to enlarge the number of latent states. However, for German-T and Korean, for example, the optimization algorithm actually chooses a smaller model than the original vanilla model.

We further inspected the behavior of the optimization algorithm for the preterminals in German-N, for which the optimal model chose (on average) a larger number of latent states. Table 5 describes this analysis. We see that in most cases, the optimization algorithm chose to decrease the number of latent states for the various preterminals, but in some cases significantly increases the number of latent states.\(^\text{13}\)

Our experiments dispel another “common wisdom” about spectral learning and training data size. It has been believed that spectral learning do not behave very well when small amounts of data are available (when compared to maximum likelihood estimation algorithms such as EM) – however we see that our results do better than the...\(^\text{14}\)

\(^{12}\)Björkelund et al. (2013) also use the MaxEnt raranker with multiple models of the Berkeley parser, and in their case also the performance after the reranking step is not always significantly better. See footnote 10 on how we create dummy head-features for our MaxEnt models.

\(^{13}\)Interestingly, most of the punctuation symbols, such as $^\text{"LRE}^\text{"}$, $^\text{\"}$. and $^\text{\"}$, drop their latent state number to a significantly lower value indicating that their interactions with other nonterminals in the tree are minimal.
Table 5: A comparison of the number of latent states for each preterminal for the German-N model, before (“b.”) running the latent state number optimization algorithm and after running it (“a.”). Note that some of the preterminals denote unary rules that were collapsed (the nonterminals in the chain are separated by |). We do not show rare preterminals with b. and a. both being 1.

Berkeley parser for several languages with small training datasets, such as Basque, Hebrew, Polish and Hungarian. The source of this common wisdom is that ML estimators tend to be statistically “efficient”: they extract more information from the data than spectral learning algorithms do. Indeed, there is no reason to believe that spectral algorithms are statistically efficient. However, it is not clear that indeed for L-PCFGs with the EM algorithm, the ML estimator is statistically efficient either. MLE is statistically efficient under specific assumptions which are not clearly satisfied with L-PCFG estimation. In addition, when the model is “incorrect,” (i.e. when the data is not sampled from L-PCFG, as we would expect from natural language treebank data), spectral algorithms could yield better results because they can mimic a higher order model. This can be understood through HMMs. When estimating an HMM of a low order with data which was generated from a higher order model, EM does quite poorly. However, if the number of latent states (and feature functions) is properly controlled with spectral algorithms, a spectral algorithm would learn a “product” HMM, where the states in the lower order model are the products of states of a higher order.\textsuperscript{14}

State-of-the-art parsers for the SPMRL datasets improve the Berkeley parser in ways which are orthogonal to the use of the basic estimation algorithm and the method for optimizing the number of latent states. They include transformations of the treebanks such as with unary rules (Björkelund et al., 2014), a more careful handling of unknown words and better use of morphological information such as decorating preterminals with such information (Björkelund et al., 2014; Szántó and Farkas, 2014), with careful feature specifications (Hall et al., 2014) and head-annotations (Crabbé, 2015), and other techniques. Some of these techniques can be applied to our case.

5 Conclusion

We demonstrated that a careful selection of the number of latent states in a latent-variable PCFG with spectral estimation has a significant effect on the parsing accuracy of the L-PCFG. We described a search procedure to do this kind of optimization, and described parsing results for eight languages (with nine datasets). Our results demonstrate that when comparing the expectation-maximization with coarse-to-fine techniques to our spectral algorithm with latent state optimization, spectral learning performs better on six of the datasets. Our results are comparable to other state-of-the-art results for these languages. Using a diverse set of models to parse these datasets further improves the results.

Acknowledgments

The authors would like to thank David McClosky for his help with running the BLLIP parser and his comments on the paper and also the three anonymous reviewers for their helpful comments. We also thank Eugene Charniak, DK Choe and Geoff Gordon for useful discussions. Finally, thanks to Djamé Seddah for providing us with the SPMRL datasets and to Thomas Müller and Anders Björkelund for providing us the MarMot models. This research was supported by an EPSRC grant (EP/L02411X/1) and an EU H2020 grant (688139/H2020-ICT-2015; SUMMA).

---

\textsuperscript{14}For example, a trigram HMM can be reduced to a bigram HMM where the states are products of the original trigram HMM.
References


David Hall, Greg Durrett, and Dan Klein. 2014. Less grammar, more features. In Proceedings of ACL.


Zsolt Szántó and Richárd Farkas. 2014. Special techniques for constituent parsing of morphologically rich languages. In Proceedings of EACL.