The Infinite Tree

Finkel et al.

presented by Vazheh Moussavi
Overview

- Introduction
- Finite Trees
- Dirichlet Process & HDP
- Infinite Trees
- Inference
- Experiments
- Conclusion
want to model syntactic dependency tree:

states: “tags”, or word categories (noun, plural noun, etc.)
observations: words
structure: tag-pair dependencies (syntax)
problem: word categories are too coarse, don’t give enough discriminative power for automatic parsers
fix #1: give tags their actual lexical form
fix #2: manually split tagset
fix #3: learn splits with heuristics
want to model syntactic dependency tree:

states: “tags”, or word categories (noun, plural noun, etc.)
observations: words
structure: tag-pair dependencies (syntax)
problem: word categories are too coarse, don’t give enough discriminative power for automatic parsers

fix #1: give tags their actual lexical form
fix #2: manually split tagset
fix #3: learn splits with heuristics
want to model syntactic dependency tree:

states: “tags”, or word categories (noun, plural noun, etc.)
observations: words
structure: tag-pair dependencies (syntax)
problem: word categories are too coarse, don’t give enough discriminative power for automatic parsers
fix #1: give tags their actual lexical form
fix #2: manually split tagset
fix #3: learn splits with heuristics
want to model syntactic dependency tree:

- states: “tags”, or word categories (noun, plural noun, etc.)
- observations: words
- structure: tag-pair dependencies (syntax)
- problem: word categories are too coarse, don’t give enough discriminative power for automatic parsers
- fix #1: give tags their actual lexical form
- fix #2: manually split tagset
- fix #3: learn splits with heuristics
Introduction/Motivation

- want to model syntactic dependency tree:

```
DT NN IN DT NN VBD PRPS NN TO VB NN EOS
The man in the corner taught his dachshund to play golf EOS
```

- states: “tags”, or word categories (noun, plural noun, etc.)
- observations: words
- structure: tag-pair dependencies (syntax)
- problem: word categories are too coarse, don’t give enough discriminative power for automatic parsers
- fix #1: give tags their actual lexical form
- fix #2: manually split tagset
- fix #3: learn splits with heuristics
“Infinite Tree”: recursive branching structure, potentially infinite states

- three different children dependency forms

- some notation:

  \[ t : \text{tree and root node} \]
  \[ c(t) : \text{list of } t\text{'s children} \]
  \[ c_i(t) : i^{th} \text{ child of } t \]
  \[ p(t) : \text{parent of } t \]
  \[ z_t \in \{1, \ldots, C\} : \text{state of } t \]
  \[ x_t : \text{observation (word) at } t \]
(conjugate) dirichlet prior $H$ on observation parameters:
\[ \phi_k \mid H \sim H \]

emissions parameterized by $F$:
\[ x_t \mid z_t \sim F(\phi_{z_t}) \]

state $z'_t$ for child of $t$ depends on $z_t$, multinomial:
\[ z'_t \mid z_t \sim \text{Multinomial}(\pi_{z_t}) \]

uniform prior on multinomial state parameters:
\[ \pi_k \mid \rho \sim \text{Dir}(\rho, \ldots, \rho) \]
Finite Trees, cont.

- **Independent Children**: (given parent)
  \[ P_{tr}(t) = P(x_t \mid z_t) \prod_{t' \in c(t)} P(z'_t \mid z_t) P_{tr}(t') \]

- **Simultaneous Children**: (no independence assumed at all)
  \[ P_{tr}(t) = P(x_t \mid z_t) P\left((z'_t)_{t' \in c(t)} \mid z_t\right) \prod_{t' \in c(t)} P_{tr}(t') \]

- \( c_t \sim \lambda_k \) (multinomial, learned)

- **Markov Children**:
  \[ P_{tr}(t) = P(x_t \mid z_t) \prod_{i=1}^{c(t)} P(z_{ci}(t) \mid z_{ci-1}(t), z_t) P_{tr}(t') \]
Finite Trees, cont.

- Independent Children: (given parent)
  \[ P_{tr}(t) = P(x_t \mid z_t) \prod_{t' \in c(t)} P(z'_{t'} \mid z_t) P_{tr}(t') \]

- Simultaneous Children: (no independence assumed at all)
  \[ P_{tr}(t) = P(x_t \mid z_t) P\left(\left( z'_{t'} \right)_{t' \in c(t)} \mid z_t \right) \prod_{t' \in c(t)} P_{tr}(t') \]

- \( c_t \sim \lambda_k \) (multinomial, learned)

- Markov Children:
  \[ P_{tr}(t) = P(x_t \mid z_t) \prod_{i=1}^{\mid c(t) \mid} P(z_{ci(t)} \mid z_{ci-1(t)}, z_t) P_{tr}(t') \]
Finite Trees, cont.

- **Independent Children: (given parent)**
  \[
P_{tr}(t) = P(x_t \mid z_t) \prod_{t' \in c(t)} P(z'_t \mid z_t) P_{tr}(t')
  \]

- **Simultaneous Children: (no independence assumed at all)**
  \[
P_{tr}(t) = P(x_t \mid z_t) P\left((z'_t)_{t' \in c(t)} \mid z_t\right) \prod_{t' \in c(t)} P_{tr}(t')
  \]

- \(c_t \sim \lambda_k\) (multinomial, learned)

- **Markov Children:**
  \[
P_{tr}(t) = P(x_t \mid z_t) \prod_{i=1}^{|c(t)|} P\left(z_{ci(t)} \mid z_{ci-1(t)}, z_t\right) P_{tr}(t')
  \]
Dirichlet Process & HDP

(a) \( \alpha_0 \)
\[ \pi \]
\[ z_i \]
\[ x_i \]
\[ N \]
\[ \pi | \alpha_0 \sim \text{GEM}(\alpha_0) \]
\[ \phi_k | H \sim H \]
\[ z_i | \pi \sim \pi \]
\[ x_i | z_i, \phi \sim F(\phi_{z_i}) \]

(b) \( \alpha_0 \)
\[ \beta \]
\[ \gamma \]
\[ H \]
\[ \phi_k \]
\[ \phi_{z_i} \]
\[ z_i \]
\[ x_i \]
\[ N \]
Infinite Trees

- mostly same, now have dirichlet process as prior
Independent Children:
mostly same, now infinite states, $\pi_k \sim DP(\alpha_0, \beta)$

HDP-HMM is same, but with only one child

Simultaneous Children:
sparisity exacerbated even further

$$P\left(\left(z'_{t} \mid t' \in c(t), \pi\right) = \prod_{t' \in c(t)} P(z'_{t} \mid \pi) = \prod_{t' \in c(t)} \pi_{z'_{t}}\right)$$

Markov Children:
same, now $\pi_{ki} \sim DP(\alpha_0, \beta)$ (from $i$ to $k$)
Independent Children:
mostly same, now infinite states, \( \pi_k \sim DP(\alpha_0, \beta) \)
HDP-HMM is same, but with only one child
Simultaneous Children:
sparsity exacerbated even further

\[
P \left( (z'_{t'})_{t' \in c(t)} \mid \pi \right) = \prod_{t' \in c(t)} P(z'_{t'} \mid \pi) = \prod_{t' \in c(t)} \pi_{z'_{t'}}
\]

\( \lambda_k \mid \zeta, L_k \sim DP(\zeta, L_k) \), \( L_k \) is a deterministic function of \( \pi_k \), acts as base measure
Markov Children:
same, now \( \pi_{ki} \sim DP(\alpha_0, \beta) \) (from \( i \) to \( k \))
Infinite Trees, cont.

- Independent Children:
  - mostly same, now infinite states, \( \pi_k \sim DP(\alpha_0, \beta) \)
  - HDP-HMM is same, but with only one child

- Simultaneous Children:
  - sparsity exacerbated even further

\[
P \left( (z_{t'})_{t' \in c(t)} \mid \pi \right) = \prod_{t' \in c(t)} P(z_{t'} \mid \pi) = \prod_{t' \in c(t)} \pi_{z_{t'}}
\]

- Markov Children:
  - \( \lambda_k \mid \zeta, L_k \sim DP(\zeta, L_k) \), \( L_k \) is a deterministic function of \( \pi_k \), acts as base measure

- same, now \( \pi_{ki} \sim DP(\alpha_0, \beta) \) (from \( i \) to \( k \))
Inference

- sampler: (1) sample state assignments $z$, (2) sample counts $m$, (3) sample global stick $\beta$
- $m_{jk}$: number of elements from $\pi_k$ corresponding to $\beta_j$
- $n_{jk}$: number of observations with state $k$ and parent state $j$
- marginal counts represented with dot ($\cdot$)
sampling \( z \)

\[
P(z_t = k \mid z^{-t}, \beta) \propto P\left(z_t = k, (z_{t'})_{t' \in s(t)} \mid z_p(t)\right) \\
\cdot P\left((z_{t'})_{t' \in s(t)} \mid z_t = k\right) \cdot f_{k}^{-x_t}(x_t)
\]

\[
f_{k}^{-x(t)}(x_t) = \frac{n_{x_{t}k} + \rho}{n_k + N\rho}
\]
Inference, cont., 2

\[
\text{sample } \beta: \\
(\beta_1, \ldots, \beta_K, \beta_u) \sim \text{Dir} \left( m_1, \ldots, m_K, \alpha_0 \right)
\]
Experiments

- only have structure, not tags:
Experiments, cont.

- results: learning and splitting tags

<table>
<thead>
<tr>
<th>Model</th>
<th>( \rho )</th>
<th># Classes</th>
<th>Acc.</th>
<th>MI</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indep.</td>
<td>0.01</td>
<td>943</td>
<td>67.89</td>
<td>2.00</td>
<td>48.29</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>1744</td>
<td>73.61</td>
<td>2.23</td>
<td>40.80</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>2437</td>
<td>74.64</td>
<td>2.27</td>
<td>39.47</td>
</tr>
<tr>
<td>Simul.</td>
<td>0.01</td>
<td>183</td>
<td>21.36</td>
<td>0.31</td>
<td>21.57</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>430</td>
<td>15.77</td>
<td>0.09</td>
<td>13.80</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>549</td>
<td>16.68</td>
<td>0.12</td>
<td>14.29</td>
</tr>
<tr>
<td>Markov</td>
<td>0.01</td>
<td>613</td>
<td>68.53</td>
<td>2.12</td>
<td>49.82</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>894</td>
<td>75.34</td>
<td>2.31</td>
<td>48.73</td>
</tr>
</tbody>
</table>

Table 1: Results of part unsupervised POS tagging on the different models, using a greedy accuracy measure.

Table 2: Results of untyped, directed dependency parsing, where the POS tags in the training data have been split according to the various models. At test time, the POS tagging and parsing are done simultaneously by the parser.