A Survey of Unsupervised Dependency Parsing

Wenjuan Han, Yong Jiang, Hwee Tou Ng, Kewei Tu
• Definition
• Generative Approaches
• Discriminative Approaches
• Recent Trends
• Further Direction
Outline

• Definition
  • Generative Approaches
  • Discriminative Approaches
  • Recent Trends
  • Further Direction
A dependency parse is a tree where
- The nodes are the words in a sentence
- The links between words represent their dependency relations
Unsupervised Dependency Parsing

Supervised Dependency Parsing

Rely on a training corpus of sentences annotated with parses (treebank)
Unsupervised Dependency Parsing

Obtained a dependency parser without using annotated sentences
Unsupervised Dependency Parsing

- Treebank may not be available for a new language or a new domain.
  - Manual annotation is labor intensive and requires linguistic knowledge and detailed guidelines.
- Unsupervised techniques can be useful in semi-supervised learning.
- Unsupervised parsing inspires/verifies cognitive research of human language acquisition.
- Grammars and parsing can be applied to other types of data. For some types of data, it is impossible to construct a treebank.
Unsupervised Dependency Parsing

Typical Pipeline of Unsupervised Dependency Parsing

I swam yesterday
Shanghai is beautiful
Empirical Method in NLP

PRP VB NN
NNP VB JJ
JJ NN IN NN

PRP VB NN
NNP VB JJ
JJ NN IN NN

Typical Pipeline of Unsupervised Dependency Parsing
Outline

- Definition
- Evaluation
- Generative Approaches
- Discriminative Approaches
- Recent Trends
- Further Direction
Outline

• Definition
• Evaluation
• Generative Approaches
• Discriminative Approaches
• Recent Trends
• Further Direction
A generative parser models: $P(\text{parse}, \text{sentence})$
A generative parser models: $P(\text{parse, sentence})$

Two steps to enable efficient inference and learning:

1. Making conditional independence assumptions (e.g., the context-free assumption)
2. Decompose the joint probability into a product of component probabilities or scores
A generative parser models: $P(\text{parse}, \text{sentence})$

Objective:

$$L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)$$

$$P(x; \Theta) = \sum_{z \in Z(x)} P(x, z; \Theta)$$
A generative parser models: \( P(\text{parse}, \text{sentence}) \)

Objective:

\[
L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)
\]

\[
P(x; \Theta) = \sum_{z \in Z(x)} P(x, z; \Theta)
\]

Priors and regularization terms are often added into the objective function to incorporate various inductive biases.
Learning

A generative parser models: $P(\text{parse, sentence})$

Objective:

$$L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)$$

$$P(x; \Theta) = \sum_{z \in Z(x)} P(x, z; \Theta)$$

Learning: Expectation-Maximization algorithm

- **E-step**: Parse the training sentences using the current grammar
- **M-step**: Update the grammar rule probability to maximize expected log likelihood of the parses ($z$) and sentences ($x$).
Learning

A generative parser models: $P(\text{parse, sentence})$

Objective:

$$L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)$$

$$P(x; \Theta) = \sum_{z \in Z(x)} P(x, z; \Theta)$$

Learning: Expectation-Maximization algorithm

- E-step: Parse the training sentences using the current grammar
- M-step: Update the grammar rule probability to maximize expected log likelihood of the parses ($z$) and sentences ($x$).

Softmax EM / Viterbi EM

Repeat until convergence
**Pros and Cons**

**Pros**
- Straightforward to incorporate inductive biases and features
- Easy training via EM

**Cons**
- Limited expressive power because of strong independence assumptions
Outline

- Definition
- Evaluation
- Generative Approaches
- Discriminative Approaches
- Recent Trends
- Further Direction
A discriminative parser models: \( P(\text{parse} \mid \text{sentence}) \)
Objective

A discriminative parser models: $P(\text{parse} \mid \text{sentence})$

Objective:
- Autoencoder-Based

$L(\Theta) = \sum_{i=1}^{N} \log P(\hat{x}^{(i)} \mid x^{(i)}; \Theta)$

- Variational Autoencoder-Based

$L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)$
Learning

A discriminative parser models: $P(\text{parse} \mid \text{sentence})$

Objective:
- Autoencoder-Based

$$L(\Theta) = \sum_{i=1}^{N} \log P(\hat{x}^{(i)} \mid x^{(i)}; \Theta)$$

- Variational Autoencoder-Based

$$L(\Theta) = \sum_{i=1}^{N} \log P(x^{(i)}; \Theta)$$

Learning: Back-propagation
Pros

- Accessing global features from the whole input sentence
- Expressive power

Cons

- Often more complicated and do not admit tractable exact inference
Generative Approaches

Performance Competition

Discriminative Approaches
Dependency Accuracy on WSJ10 Testset
(Training with WSJ10, no lexicalization)
### Detailed Performance

<table>
<thead>
<tr>
<th>Methods</th>
<th>≤ 10</th>
<th>ALL</th>
<th>Generative Approaches (cont’d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klein and Manning (2004)</td>
<td>46.2</td>
<td>34.9</td>
<td>Spitkovsky et al. (2011a)</td>
</tr>
<tr>
<td>Cohen et al. (2008)</td>
<td>59.4</td>
<td>40.5</td>
<td>Gimpel and Smith (2012)</td>
</tr>
<tr>
<td>Cohen and Smith (2009)</td>
<td>61.3</td>
<td>41.4</td>
<td>Tu and Honavar (2012)</td>
</tr>
<tr>
<td>Headden III et al. (2009)</td>
<td>68.8</td>
<td>-</td>
<td>Bisk and Hockenmaier (2012)</td>
</tr>
<tr>
<td>Spitkovsky et al. (2010a)</td>
<td>56.2</td>
<td>44.1</td>
<td>Spitkovsky et al. (2013)</td>
</tr>
<tr>
<td>Berg-Kirkpatrick et al. (2010)</td>
<td>63.0</td>
<td>-</td>
<td>Jiang et al. (2016)</td>
</tr>
<tr>
<td>Gillenwater et al. (2010)</td>
<td>64.3</td>
<td>53.3</td>
<td>Han et al. (2017)</td>
</tr>
<tr>
<td>Spitkovsky et al. (2010b)</td>
<td>65.3</td>
<td>47.9</td>
<td>He et al. (2018)*</td>
</tr>
<tr>
<td>Naseem et al. (2010)</td>
<td>71.9</td>
<td>-</td>
<td>Le and Zuidema (2015) †</td>
</tr>
<tr>
<td>Blunsom and Cohn (2010)</td>
<td>67.7</td>
<td>55.7</td>
<td>Cai et al. (2017)</td>
</tr>
<tr>
<td>Spitkovsky et al. (2011c)</td>
<td>-</td>
<td>55.6</td>
<td>Li et al. (2019)</td>
</tr>
<tr>
<td>Spitkovsky et al. (2011b)</td>
<td>69.5</td>
<td>58.4</td>
<td>Han et al. (2019a)</td>
</tr>
</tbody>
</table>

*Reported directed dependency accuracies on section 23 of the WSJ corpus, evaluated on sentences of length ≤ 10 and all lengths. *: without golden POS tags. †: with more training data in addition to WSJ.
Outline

• Definition
• Evaluation
• Generative Approaches
• Discriminative Approaches
• Recent Trends
• Further Direction
Recent Trends

- Combined Approaches
- Neural Parameterization
- Lexicalization
- Big Data
- Unsupervised Multilingual Parsing
Outline

• Definition
• Evaluation
• Generative Approaches
• Discriminative Approaches
• Recent Trends
• Further Direction
Further Directions

- Syntactic Information in Pretrained Language Modeling
- Inspiration for Other Tasks
- Interpretability
Thank you

Wenjuan Han
2020/12