A Survey of Unsupervised Dependency Parsing

Wenjuan Han, Yong Jiang, Hwee Tou Ng, Kewei Tu

NUS National University of Singapore

上海科技大学 ShanghaiTech University

達摩院 ALIBABA DAMO ACADEMY

Outline

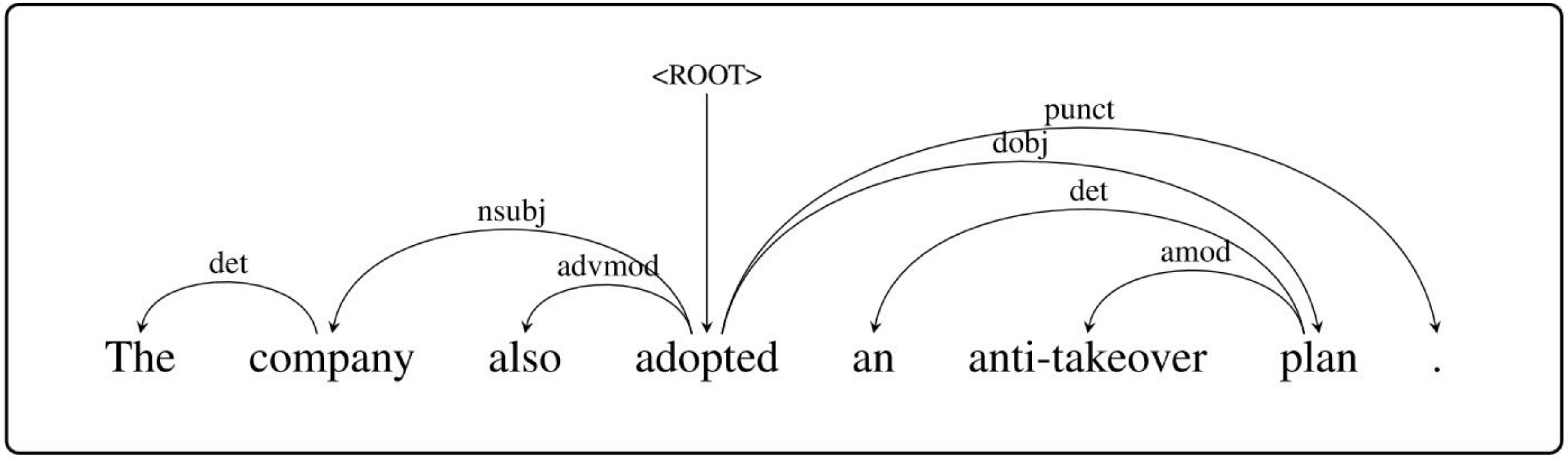
- Definition
- Generative Approaches
- Discriminative Approaches
- Recent Trends
- Further Direction

Outline

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

Dependency Parsing

- 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

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

Outline

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


Discriminative
Approaches
Generative
Approaches

Outline

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

Model

A generative parser models: P(parse, sentence)

Model

A generative parser models: P(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

Objective

A generative parser models: P(parse, sentence)

Objective:

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

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

Objective

A generative parser models: P(parse, sentence)

Objective:

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

Vanilla

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

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

Learning

A generative parser models: P(parse, sentence)

Objective:

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

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

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).

Repeat until convergence

Learning

A generative parser models: P(parse, sentence)

Objective:

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

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

Standard

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).

Repeat until convergence

Softmax EM / Viterbi EM

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

Model

A discriminative parser models: P(parse | sentence)

Objective

A discriminative parser models: P(parse | sentence)

Objective:

- Autoencoder-Based

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

- Variational Autoencoder-Based

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

Learning

A discriminative parser models: P(parse | sentence)

Objective:

- Autoencoder-Based

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

- Variational Autoencoder-Based

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

Learning: Back-propagation

Pros and Cons

Pros

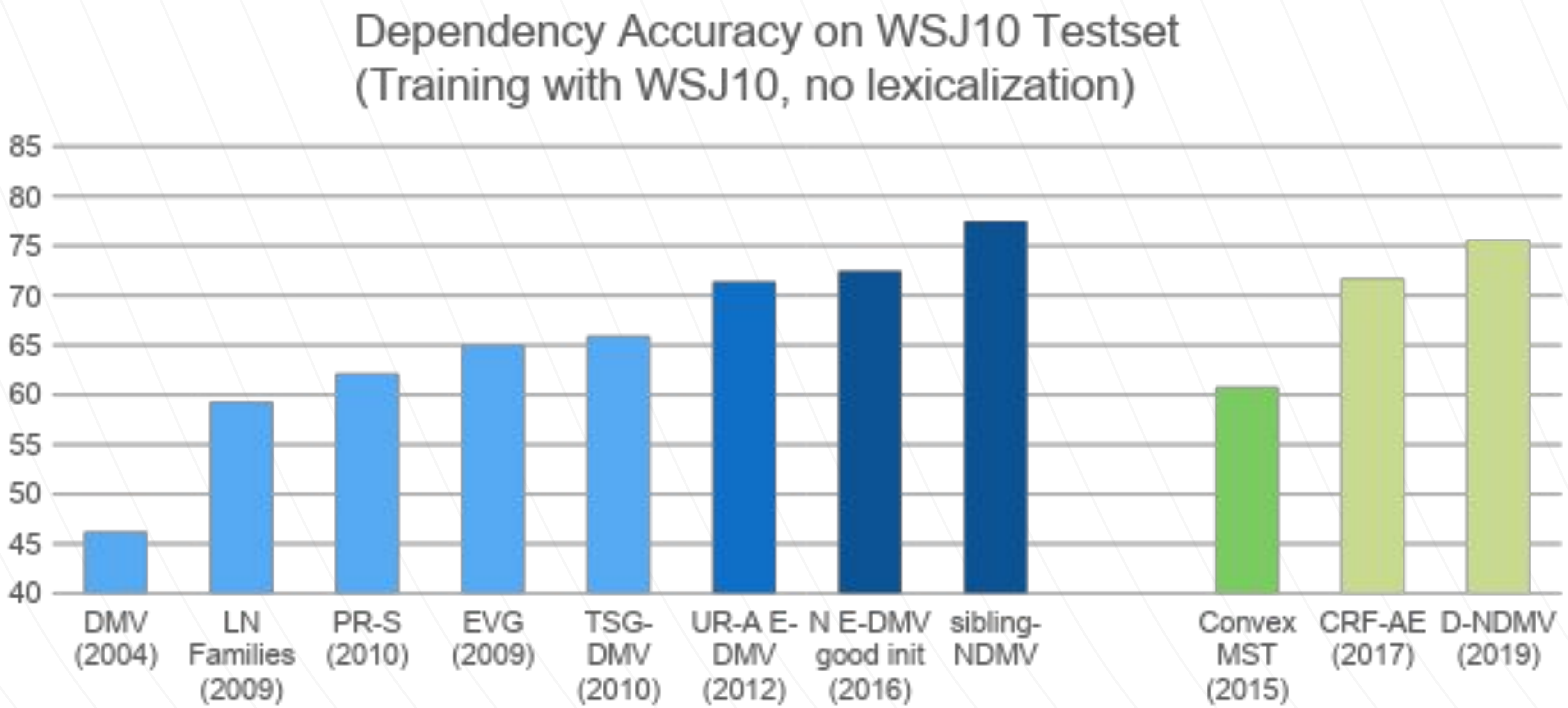
- Accessing global features from the whole input sentence
- Expressive power

Cons

- Often more complicated and do not admit tractable exact inference

Performance Competition

Performance



Detailed Performance

METHODS	≤ 10	ALL	Generative Approaches (cont'd)		
Generative Approaches			Spitkovsky et al. (2011a)	-	59.1
Klein and Manning (2004)	46.2	34.9	Gimpel and Smith (2012)	64.3	53.1
Cohen et al. (2008)	59.4	40.5	Tu and Honavar (2012)	71.4	57.0
Cohen and Smith (2009)	61.3	41.4	Bisk and Hockenmaier (2012)	71.5	53.3
Headden III et al. (2009)	68.8	-	Spitkovsky et al. (2013)	72.0	64.4
Spitkovsky et al. (2010a)	56.2	44.1	Jiang et al. (2016)	72.5	57.6
Berg-Kirkpatrick et al. (2010)	63.0	-	Han et al. (2017)	75.1	59.5
Gillenwater et al. (2010)	64.3	53.3	He et al. (2018)*	60.2	47.9
Spitkovsky et al. (2010b)	65.3	47.9	Discriminative Approaches		
Blunsom and Cohn (2010)	65.9	53.1	Daumé III (2009)	-	45.4
Naseem et al. (2010)	71.9	-	Le and Zuidema (2015) †	73.2	65.8
Blunsom and Cohn (2010)	67.7	55.7	Cai et al. (2017)	71.7	55.7
Spitkovsky et al. (2011c)	-	55.6	Li et al. (2019)	54.7	37.8
Spitkovsky et al. (2011b)	69.5	58.4	Han et al. (2019a)	75.6	61.4

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

2020

QA