Unsupervised Natural Language Parsing

Kewei Tu ShanghaiTech University
Yong Jiang Alibaba DAMO Academy
Wenjuan Han National University of Singapore
Yanpeng Zhao University of Edinburgh

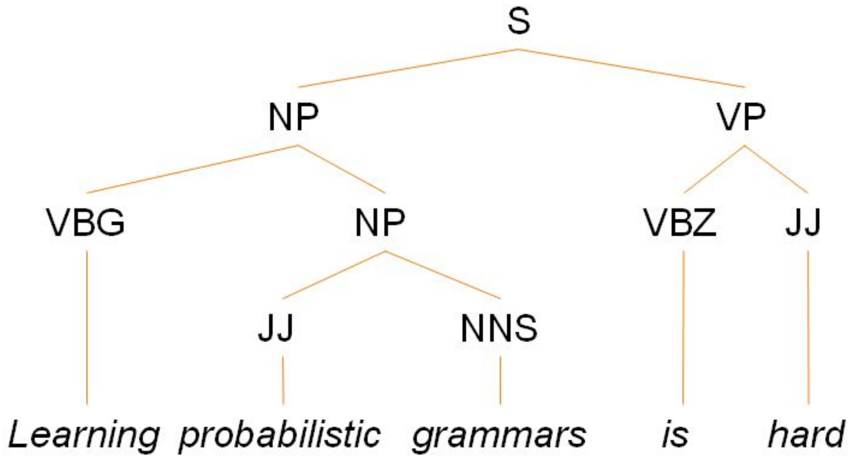
Tutorial Overview

1.	Introduction	(Kewei)
2.	Generative Approaches	(Kewei, Yong)
3.	Discriminative Approaches	(Wenjuan)
4.	Special Topics	(Yanpeng)
5.	Summary	(Kewei)

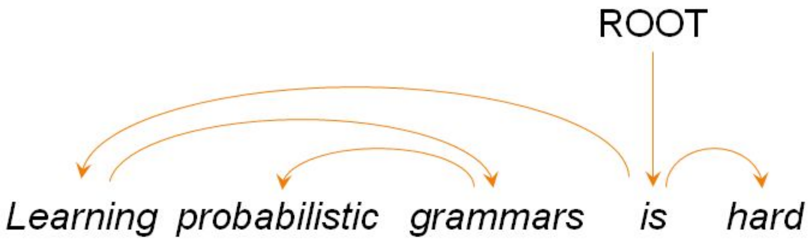
1. Introduction

Syntactic Parsing

- Goal: identifying the syntactic structure of a sentence
    - Typically a tree structure over the sequence of words

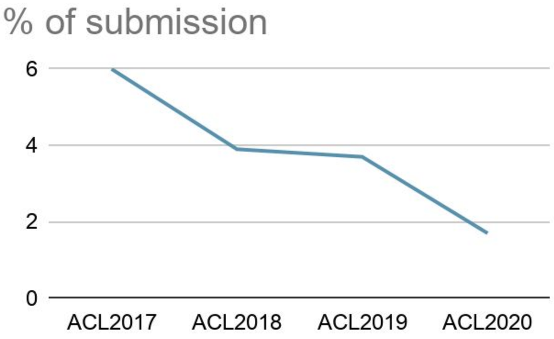


Constituency Parsing



Dependency Parsing

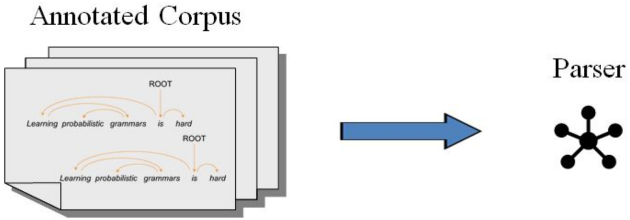
Syntactic Parsing



- In traditional NLP:
  - A key component in the NLU pipeline
- In the era of deep learning:
  - Diminishing importance...
    - Sequential models (+attention) seem to work very well.
  - ...but regains some attention in recent years
    - Ex: useful in some tasks, such as SRL (Strubell et al., 2018)
    - Ex: knowledge distillation from RNNG (a syntactic parser/LM) to BERT (Kuncoro et al., 2020)
- Our thoughts:
  - Linguistic structures are an intrinsic property of natural languages
  - We should utilize them instead of ignoring them

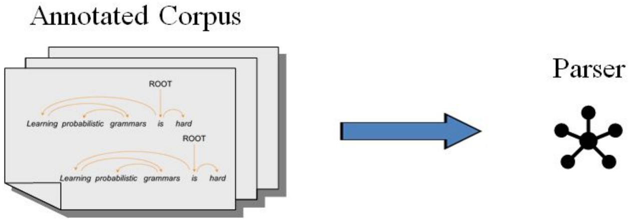
Supervised vs. Unsupervised Parsing

- Supervised parsing: learning a parser from training sentences annotated with parses (treebank)


Annotated Corpus
ROOT
Learning probabilistic grammars is hard
ROOT
Learning probabilistic grammars is hard
Parser

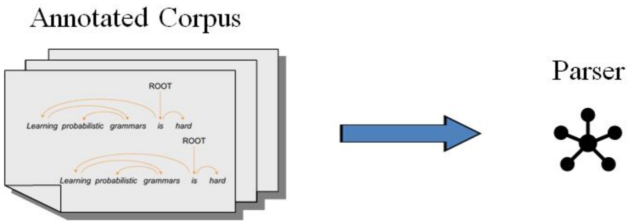
Supervised vs. Unsupervised Parsing

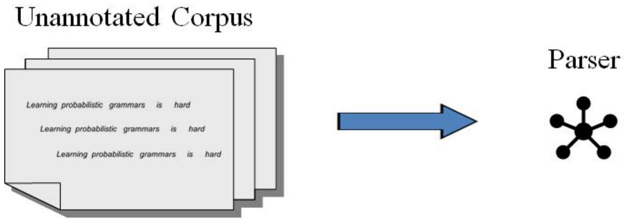
- Supervised parsing: learning a parser from training sentences annotated with parses (treebank)
    - Limitation: shortage of high-quality treebanks in low-resource languages or domains


Annotated Corpus
ROOT
Learning probabilistic grammars is hard
ROOT
Learning probabilistic grammars is hard
Parser

Supervised vs. Unsupervised Parsing

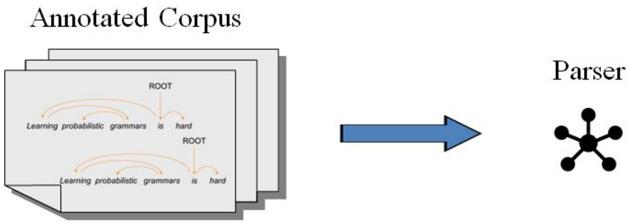
- Supervised parsing: learning a parser from training sentences annotated with parses (treebank)
  - Limitation: shortage of high-quality treebanks in low-resource languages or domains
- Unsupervised parsing: learning a parser without annotated data

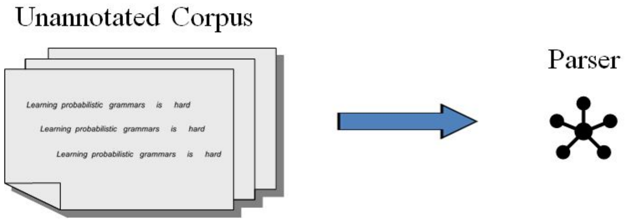

Annotated Corpus
ROOT
Learning probabilistic grammars is hard
ROOT
Learning probabilistic grammars is hard
Parser


Unannotated Corpus
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Parser

Supervised vs. Unsupervised Parsing

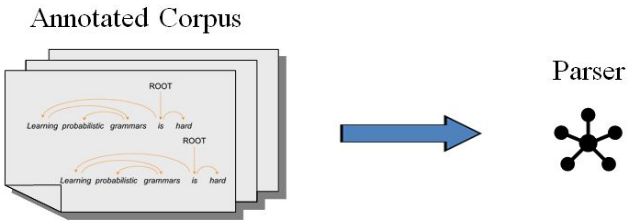
- Supervised parsing: learning a parser from training sentences annotated with parses (treebank)
  - Limitation: shortage of high-quality treebanks in low-resource languages or domains
- Unsupervised parsing: learning a parser without annotated data
  - Typical setting: learning from unannotated data

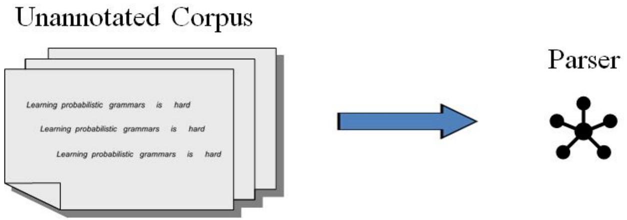

Annotated Corpus
ROOT
Learning probabilistic grammars is hard
ROOT
Learning probabilistic grammars is hard
Parser


Unannotated Corpus
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Parser

Supervised vs. Unsupervised Parsing

- Supervised parsing: learning a parser from training sentences annotated with parses (treebank)
  - Limitation: shortage of high-quality treebanks in low-resource languages or domains
- Unsupervised parsing: learning a parser without annotated data
  - Typical setting: learning from unannotated data
  - Exception exists: no data at all (Søgaard, 2012)

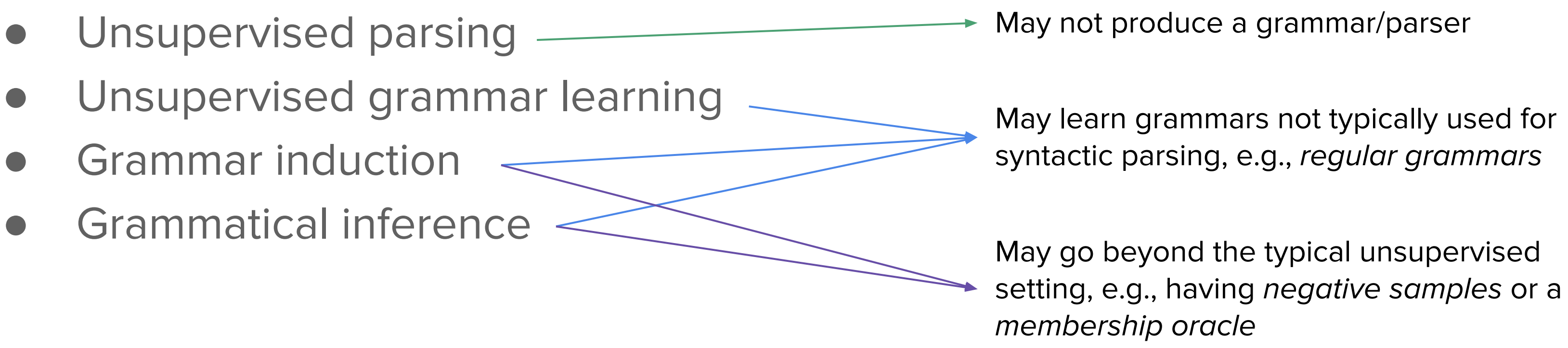

Annotated Corpus
ROOT
Learning probabilistic grammars is hard
ROOT
Learning probabilistic grammars is hard
Parser


Unannotated Corpus
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Learning probabilistic grammars is hard
Parser

Why Unsupervised?

- It requires no human annotation (good for low-resource settings)
- It can utilize (potentially unlimited) unannotated text data.
- It serves as the basis for semi-supervised, weakly supervised, and transfer learning of syntactic parsers.
    - Ex: CRFAE unsupervised (Cai et al., 2017), semi-supervised (Jia et al., 2020; Zhang & Goldwasser, 2020), cross-lingual transfer (Li & Tu, 2020)
- It is a representative task of unsupervised structured prediction.
- It inspires/verifies cognitive research of human language acquisition.
- It can be extended to data of other modalities without treebanks.
    - Ex: image parsing (Tu et al., 2013), probabilistic modeling (Poon & Domingos, 2011)

Terminology



- Unsupervised parsing
- Unsupervised grammar learning
- Grammar induction
- Grammatical inference

May not produce a grammar/parser

May learn grammars not typically used for syntactic parsing, e.g., *regular grammars*

May go beyond the typical unsupervised setting, e.g., having *negative samples* or a *membership oracle*

History

- A long history
    - Language identification in the limit (Gold, 1967)
    - Inside-outside algorithm (Baker, 1979)
- Recent surge of interest
    - [a plot of #paper over the past 3-5 years?]
    - [a figure of two trends:]
        - a general trend in deep learning towards unsupervised training or pre-training
        - an emerging trend in the NLP community towards finding or modeling linguistic structures in neural models

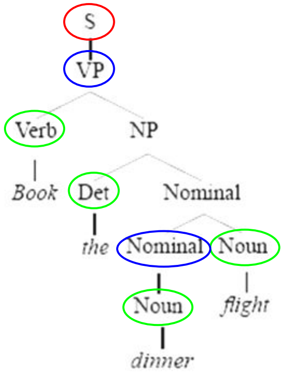
Evaluation -- typical experimental setup

- Availability of POS annotations
  - Exceptions: induced POS tags (Spitkovsky et al., 2011a; He et al., 2018), no POS (Seginer, 2007; Pate & Johnson, 2016)
- Length limit of training sentences
  - Many methods work best with a length limit of 10-15 for English
  - More recent methods are able to learn from longer sentences
- Punctuation removal
  - Punctuation marks can provide info of phrase boundaries; simply treating them as words may hurt learning (Spitkovsky et al., 2011b)

Evaluation -- metrics

Constituency parsing

- F1 score: the harmonic mean of precision & recall of constituents
  - Precision: the percentage of predicted constituents that are correct
  - Recall: the percentage of gold constituents that are predicted
- Removing trivial constituents
  - Single-word spans
  - Whole-sentence spans
  - Duplicate spans



⚠ *Many previous studies follow different practices.*

Evaluation -- metrics

Dependency parsing

- Directed dependency accuracy (DDA)
  - Percentage of correctly predicted dependencies
- Undirected dependency accuracy (UDA)
  - Percentage of correctly predicted dependencies when ignoring their directions
- Neutral edge detection (NED) (Schwartz et al., 2011)
  - Similar to UDA, but allows that the predicted parent of a token is actually the grandparent

Evaluation -- metrics

Micro-average (i.e., corpus-level score)

- Aggregating the predicted and gold constituents/dependencies from all the sentences and then calculating the score

Macro-average (i.e., sentence-level score)

- Calculating the score for each individual sentence and then take an average over all the sentences

⚠ *Many previous studies use different averaging methods.*

Evaluation -- hyperparameter tuning ⚠

- A lot of previous studies perform hyperparameter tuning with evaluation metrics (e.g., F1 or DDA) on a development corpus annotated with parse trees
- Consequences:
  - Learning is no longer purely unsupervised.
  - It calls for comparison with supervised learning on the dev set
    - This has been found to outperform unsupervised parsing (Shi et al., 2020).
- Alternative strategies
  - Perform hyperparameter tuning with metrics not based on gold parses, e.g., perplexity
  - Perform hyperparameter tuning with gold parses on one language, but fix the hyperparameter values during evaluation on other languages

Evaluation -- Are gold parses unique?

- There exist different linguistic theories resulting in different gold parses
    - Ex: Some theories choose determiners as the heads of noun phrases (Abney, 1987).
    - Since unsupervised parsing has no clue what theory it should follow, isn’t it problematic to do evaluation with specific theories embodied by available treebanks?
- Solution?
    - Ultimately, parsing is supposed to provide useful info for downstream tasks.
    - One may use performance on downstream tasks as a surrogate metric.