An Imitation Learning Approach to Unsupervised Parsing

Bowen Li,[e] Lili Mou,[a] Frank Keller[e]

e: University of Edinburgh Natural Language Processing Edinburgh NLP

a: University of Alberta

Why Unsupervised Parsing?

Engineering motivation:

- $\sim$6,000 languages in the world
- Treebanks for $\sim$70 languages (many of them small)
- Syntactic annotation
  - slow and costly
  - relying on expert linguists

We need a way of inducing syntactic knowledge

- Based on simple, crowd-sourcable sentence annotation
- E.g., natural language inference, sentiment

Why Unsupervised Parsing?

Cognitive motivation: how children learn languages?

- 18 months: start with two word utterances
- By 5 years: generate complex syntax (Brown's stages):
  - relative clauses, infinitival, gerunds, wh-phrases, passives
- No explicit supervision is provided (children don't see syntax trees)
- But they receive indirect feedback: is an utterance understood or not?

To model this, we need a way of inducing syntactic knowledge based on simple semantic labels at the sentence level

Unsupervised Parsing

Goal: learn linguistically meaningful syntax (tree structures) without treebank supervision

Approach:

- Get training signal from a secondary task:
  - Language modeling
  - Semantically oriented tasks (e.g., natural language inference, sentiment)
- Try to induce meaningful "latent" tree structures

Hard Discrete Parsers

Examples:

- RL-SPINN [Yogatama et al., 2017], Soft-Gating [Maillard et al., 2017], Gumbel-Tree-LSTM [Choi et al., 2018]

Advantages:

- Models have grounded parsing actions

Disadvantages: Not differentiable

- Reinforcement learning $\Longrightarrow$ doubly stochastic gradient descent, poor local optima, low self-agreement
- Dynamic Programming marginalization $\Longrightarrow$ high time complexity

Soft Continuous Parsers

Very recent work:

- Parsing-reading-predict network [PRPN, Shen et al., 2018]
- Ordered Neurons [ON-LSTM, Shen et al., 2019]

Advantages:

- Relaxing discrete parsing by continuous notions (e.g., structured attention) $\Longrightarrow$ easy to train by differentiation

Disadvantages:

- Inducing syntax from continuous relaxation is **not learnable**
- Parsing operations are stipulated externally by heuristics

Combine both Worlds by Imitation Learning

- Is it possible to combine both approaches?

Combine both Worlds by Imitation Learning

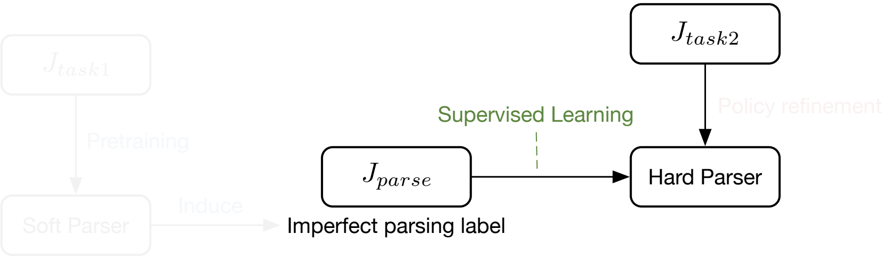
- Is it possible to combine both approaches?
- Yes! We can use imitation learning!

Combine both Worlds by Imitation Learning

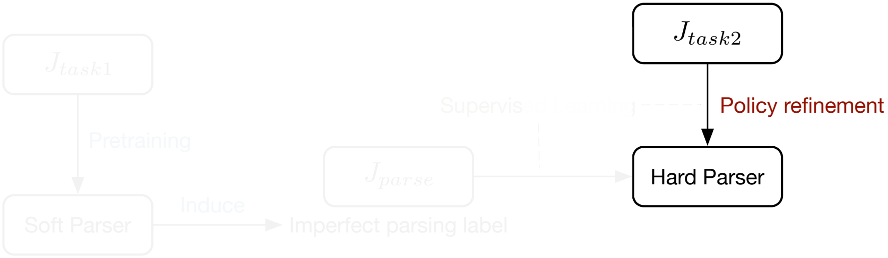
- Is it possible to combine both approaches?
- Yes! We can use imitation learning!
- Coupling soft continuous parser and hard discrete parser at the intermediate output level (parse tree)

Combine both Worlds by Imitation Learning

Combine both Worlds by Imitation Learning



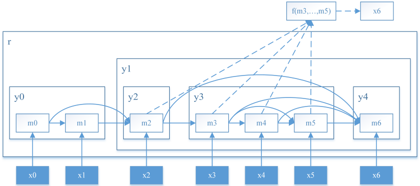
Combine both Worlds by Imitation Learning



PRPN as the Soft Parser

Parsing-reading-predict network (PRPN; [Shen et al. 2018])

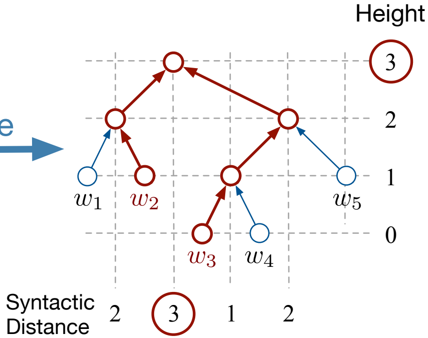
LSTM with structured attention for LM



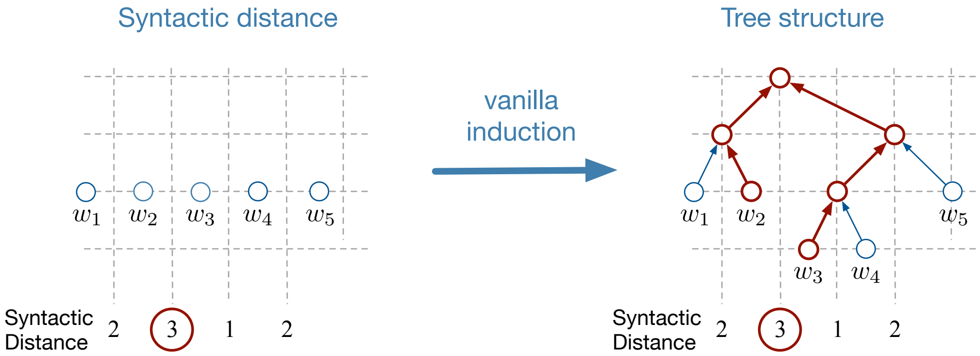
[Shen et al., 2018]

Induce

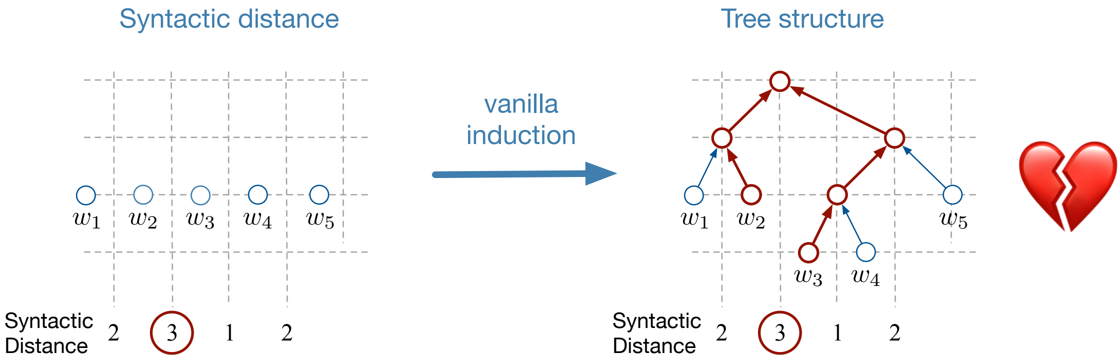
Syntactic distance



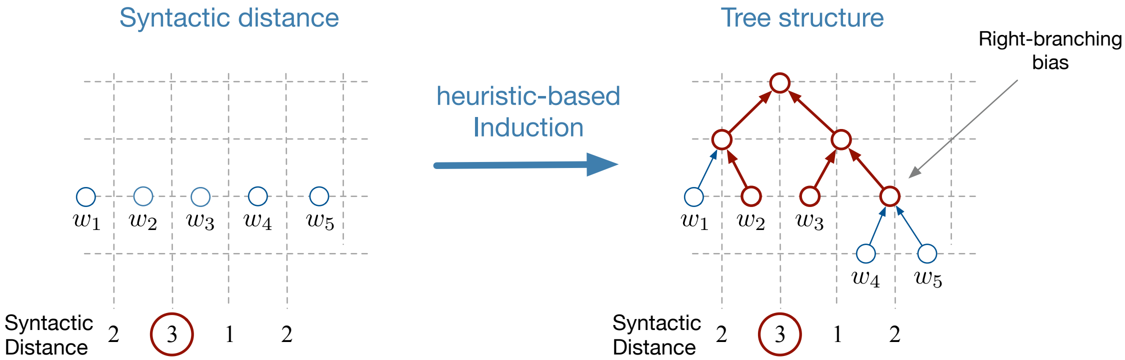
PRPN as the Soft Parser



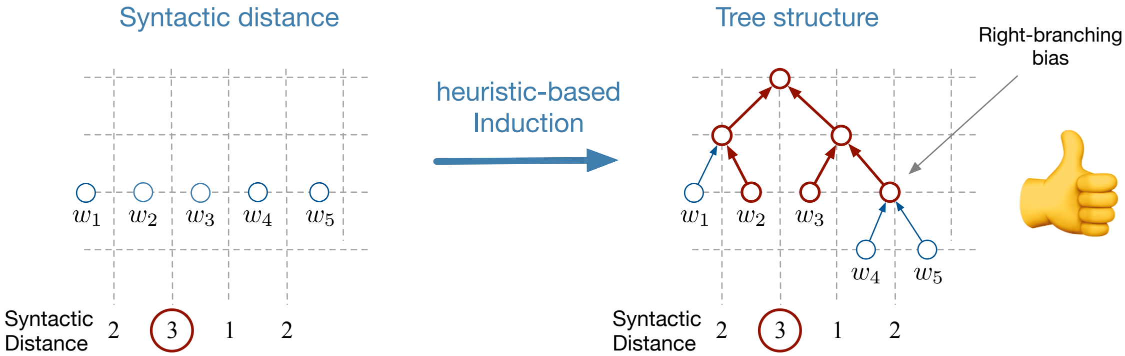
PRPN as the Soft Parser



PRPN as the Soft Parser

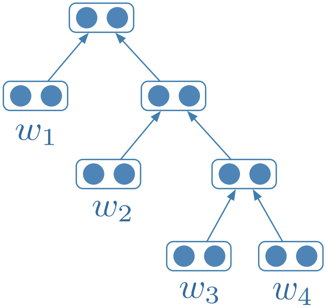


PRPN as the Soft Parser

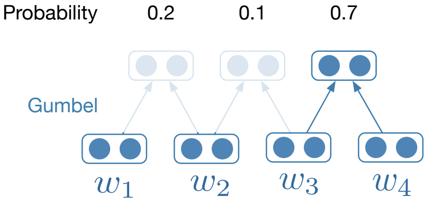


Gumbel-Tree-LSTM as the Hard Parser

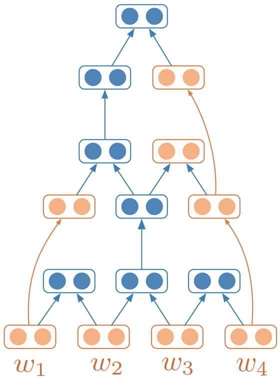
Tree-LSTM for sentence classification



Learning tree structures by Straight-Through Gumbel Softmax



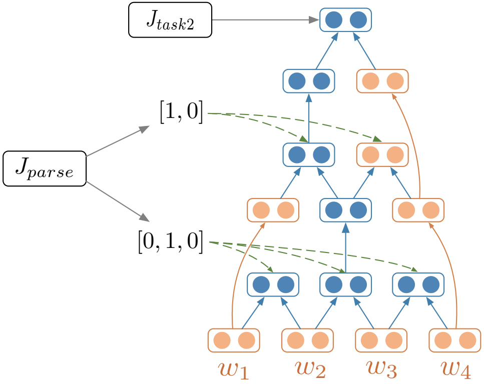
Our Approach



Our Approach

Two-stage training:

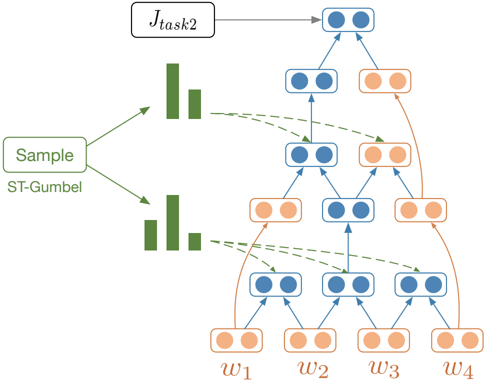
- Stage 1: step-by-step supervised learning



Our Approach

Two-stage training:

- Stage 1: step-by-step supervised learning
- Stage 2: policy refinement on NLI task



Experimental Results: Parsing Results on All-NLI

Model	Mean F	Self-agreement
Left-Branching	18.9	-
Right-Branching	18.5	-
Balanced-Tree	22.0	-
Gumbel-Tree-LSTM	21.9	56.8
PRPN	51.6	65.0

Experimental Results: Parsing Results on All-NLI

Model	Mean F	Self-agreement
Left-Branching	18.9	-
Right-Branching	18.5	-
Balanced-Tree	22.0	-
Gumbel-Tree-LSTM	21.9	56.8
PRPN	51.6	65.0
Imitation (first stage only)	52.0	**70.8**

Experimental Results: Parsing Results on All-NLI

Model	Mean F	Self-agreement
Left-Branching	18.9	-
Right-Branching	18.5	-
Balanced-Tree	22.0	-
Gumbel-Tree-LSTM	21.9	56.8
PRPN	51.6	65.0
Imitation (first stage only)	52.0	**70.8**
Imitation (two stages)	**53.7**	67.4

More settings and analysis in our paper

Relationship to Previous Studies

Do latent tree learning models identify meaningful structure in sentences?

[Williams et al., 2018]

- Our results: Yes, but we need a “good” initialization.

Relationship to Previous Studies

Do latent tree learning models identify meaningful structure in sentences? [Williams et al., 2018]

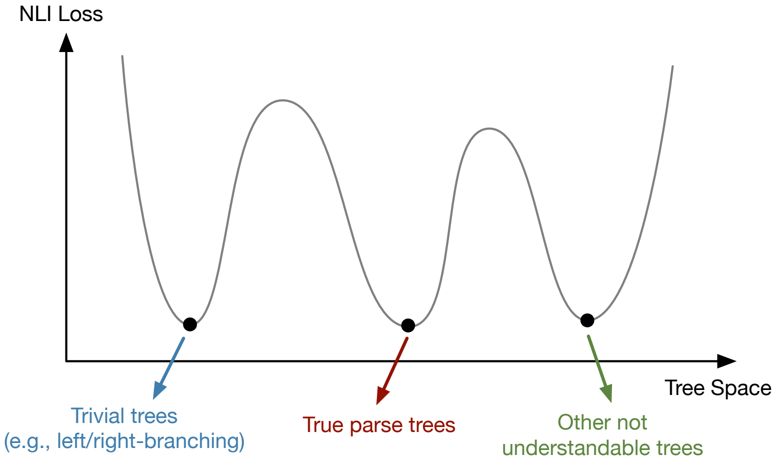
- Our results: Yes, but we need a "good" initialization.

Tree-Based Neural Sentence Modeling [Shi et al., 2018]: parse/trivial trees are roughly the same for classification performance

- Our results: same findings in terms of NLI accuracy

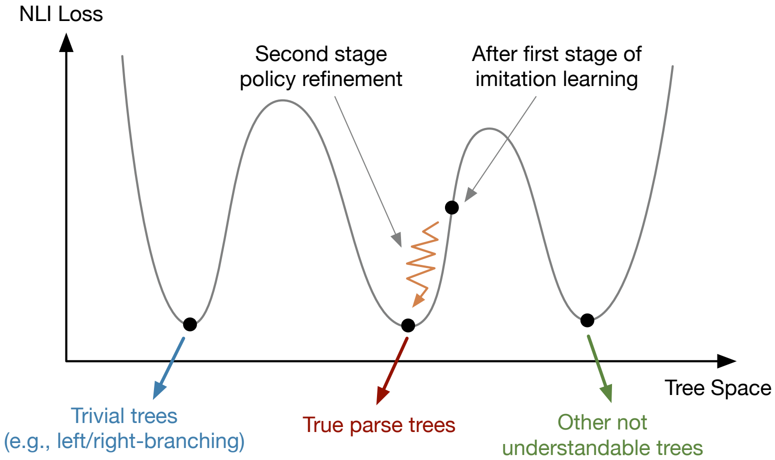
One last question

Why does NLI help unsupervised parsing?



One last question

Why does NLI help unsupervised parsing?



Conclusion

- Imitation learning for unsupervised parsing
  - A flexible way of coupling heterogeneous models on the intermediate output level
  - Other applications: semantic parsing [Mou et al., 2017], discourse parsing
- Showing the usefulness of semantic tasks for unsupervised parsing
- More research needed on tasks, models, and combinations in this direction

Thank you!

Q&A