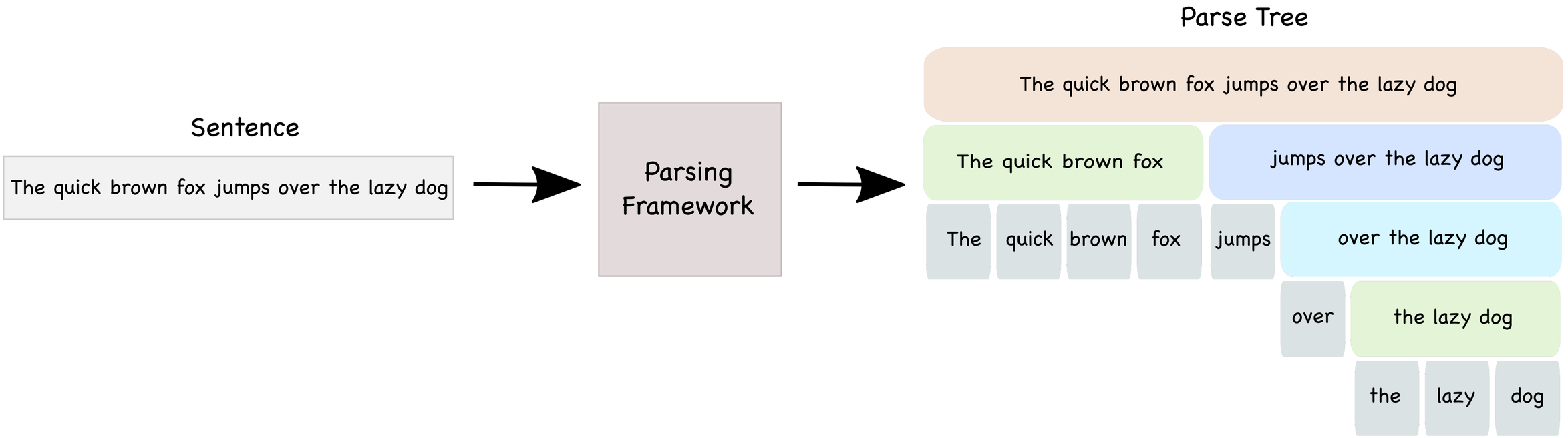
Co-training an Unsupervised Constituency Parser with Weak Supervision

Nickil Maveli and Shay Cohen

THE UNIVERSITY of EDINBURGH
informatics

Unsupervised Constituency Parsing



Goal: Induce parse trees from observed sentences alone without supervision

Motivation

- Current supervised parsers operate on a minuscule of commonly spoken languages in the world.
- The process of annotation of syntactic trees by human language experts is often associated with high-costs and is time-intensive.
- Lack of clear annotation rubrics for certain low-resource languages.
- Annotations lack ability to scale to out-of-domain data.

Previous Approaches

- **Generative:** Models the joint probability distribution P(x, z) over sentence x and parse tree z through a grammar component.
  - Constituent Context Model (CCM) Klein and Manning (2002)
  - Parsing-Reading Predict Network (PRPN) Shen et al. (2018b)
  - Ordered Neurons (ON) Shen et al. (2019)
  - Unsupervised Recurrent Neural Network Grammars (URNNG) Kim et al. (2019b)
- **Discriminative:** Models the conditional probability P(z | x) of the output parse tree z conditioned on the sentence x.
  - Deep Inside-Outside Recursive Autoencoders (DIORA) Drozdov et al. (2019)
  - Compound PCFG Kim et al. (2019)
  - S-DIORA Drozdov et al. (2020)

Datasets

- Penn Treebank (PTB)

> Nissan hopes that that will start to change this fall , with its new version of the Stanza compact sedan .

- Chinese Treebank (CTB)

> 西藏 银行 部门 积极 调整 信贷 结构 以 确保 农牧业 生产 等 重点 产业 的 投入 加大 对 工业 能源 交通 通信 等 建设 的 正常 资金 供应量

- Keyaki Treebank (KTB)

> しかし 二 度目 の 車輪 の 音 は 、 もう 彼 を 驚かさ なかっ た 。

Weak Supervision

- Branching Direction

Data	Branching
PTB	Mostly Right
CTB	Mixed
KTB	Mostly Left

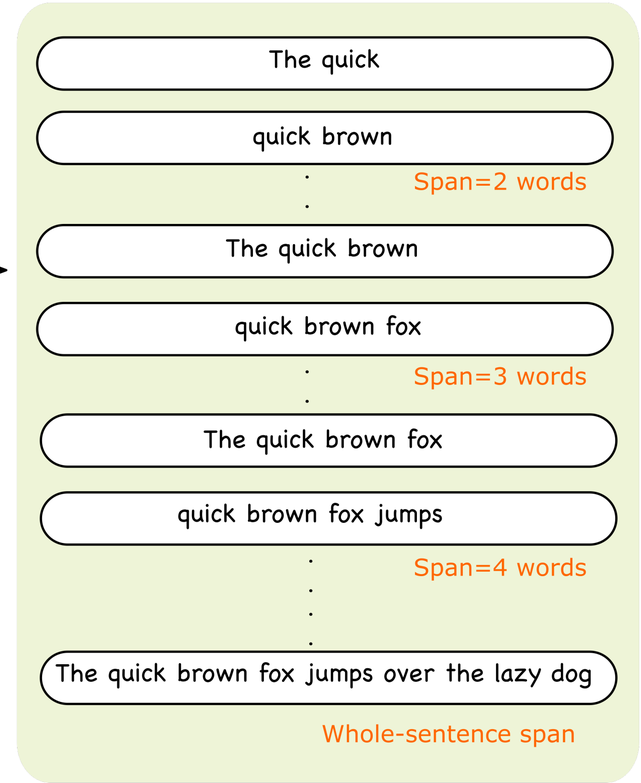
- Rule-based Heuristics

Data	Rules?
PTB	✓
CTB	✗
KTB	✗

Inside and Outside Strings

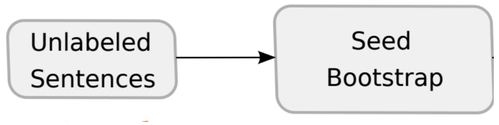
Sentence

The quick brown fox jumps over the lazy dog

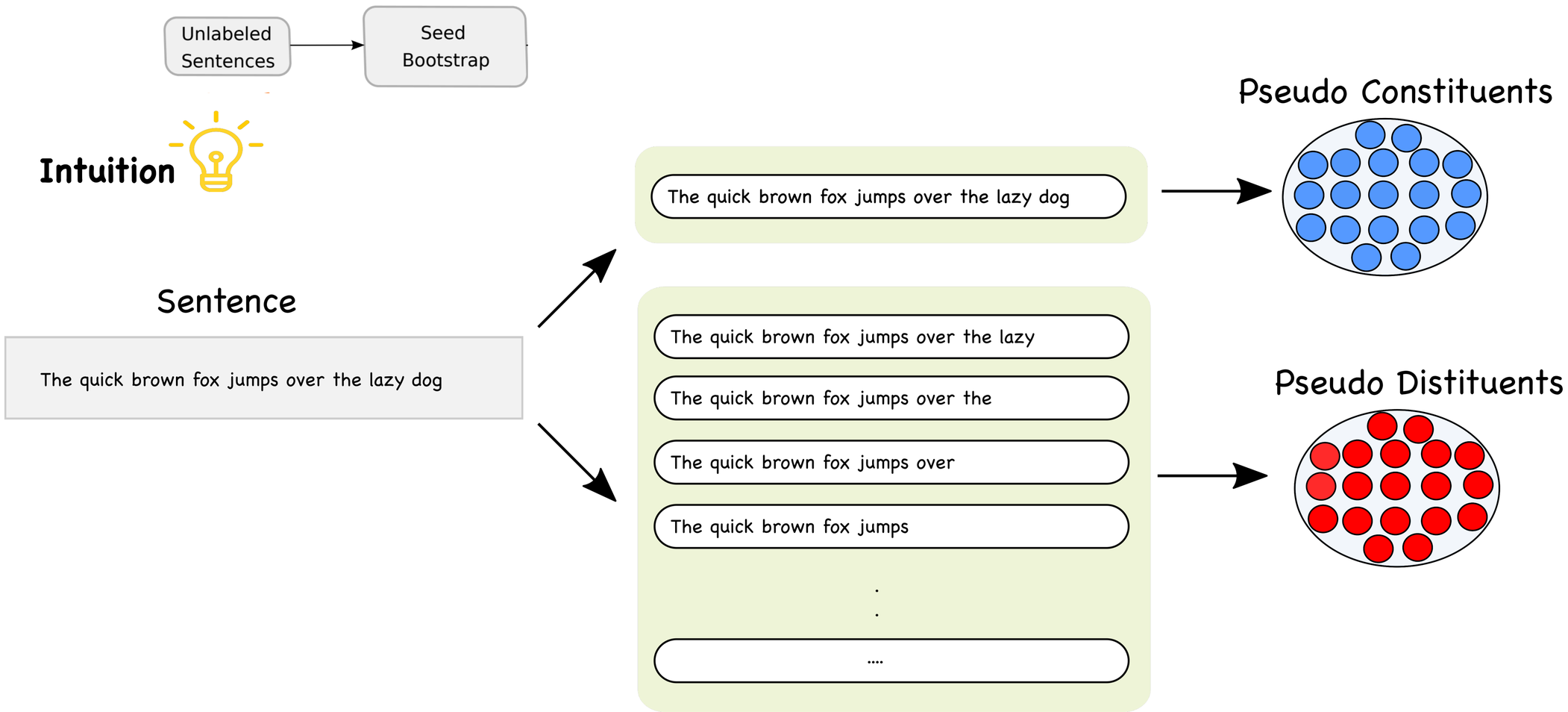


Inside string	Outside string
The quick	<BOS>, <MASK>, brown
quick brown	The, <MASK>, fox
.	.
.	.
The quick brown	<BOS>, <MASK>, fox
quick brown fox	The, <MASK>, jumps
.	.
.	.
The quick brown fox	<BOS>, <MASK>, jumps
quick brown fox jumps	The, <MASK>, over
.	.
.	.
.	.
.	.
The quick brown fox jumps over the lazy dog	<BOS>, <MASK>, <EOS>

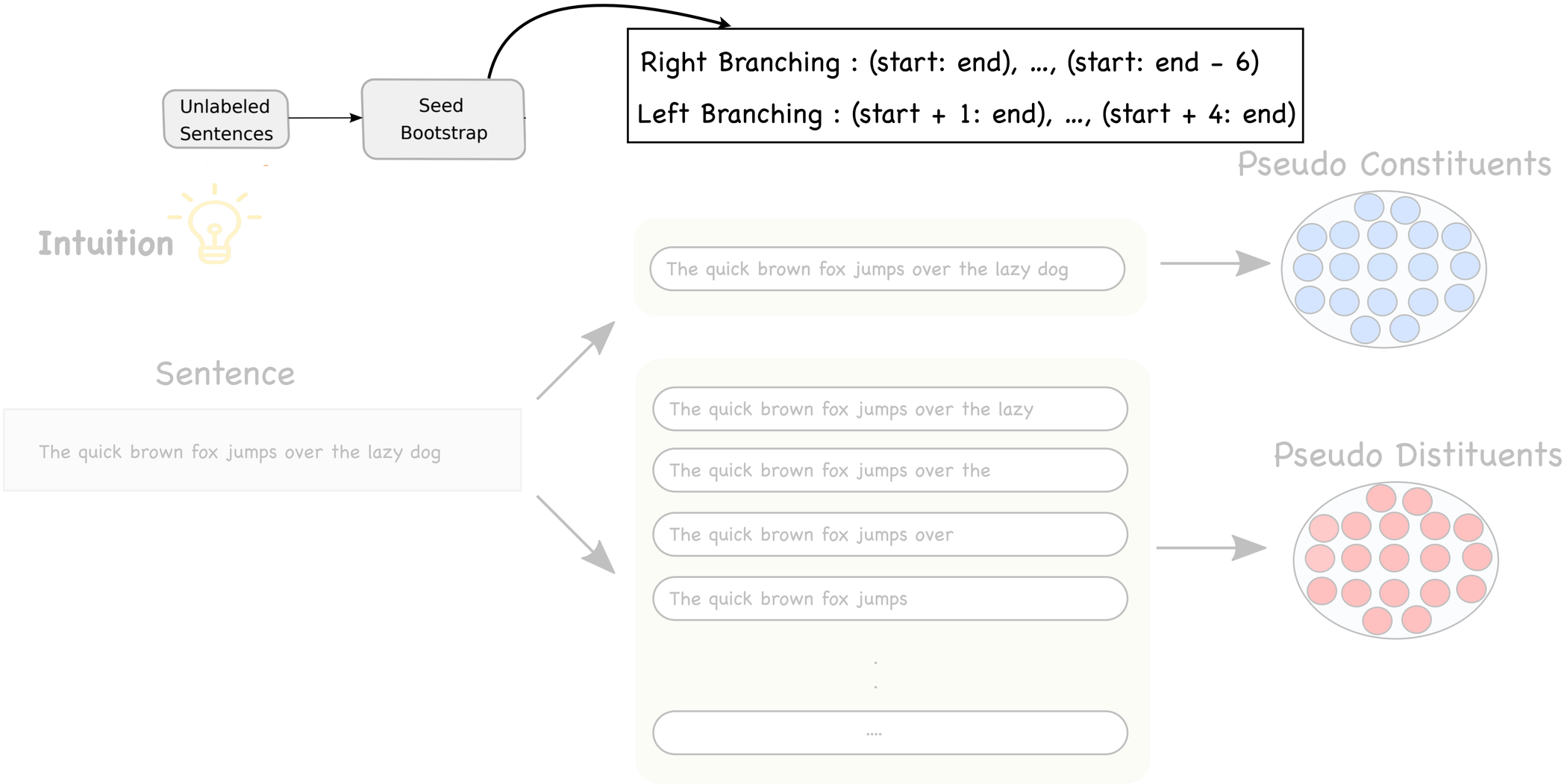
Implementation: Seed Bootstrapping



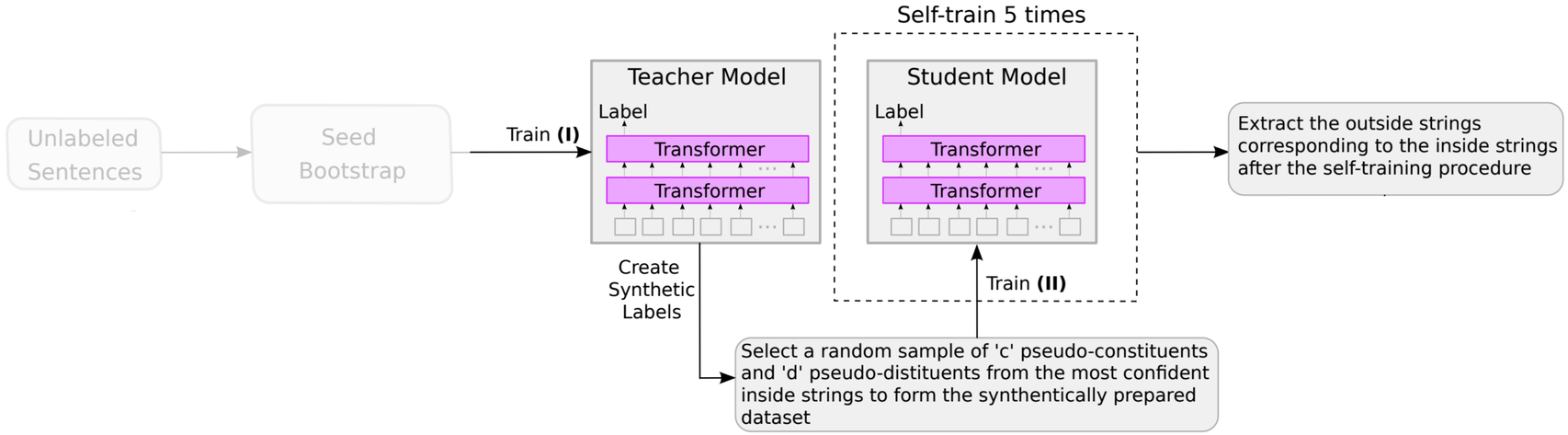
Implementation: Seed Bootstrapping



Implementation: Seed Bootstrapping



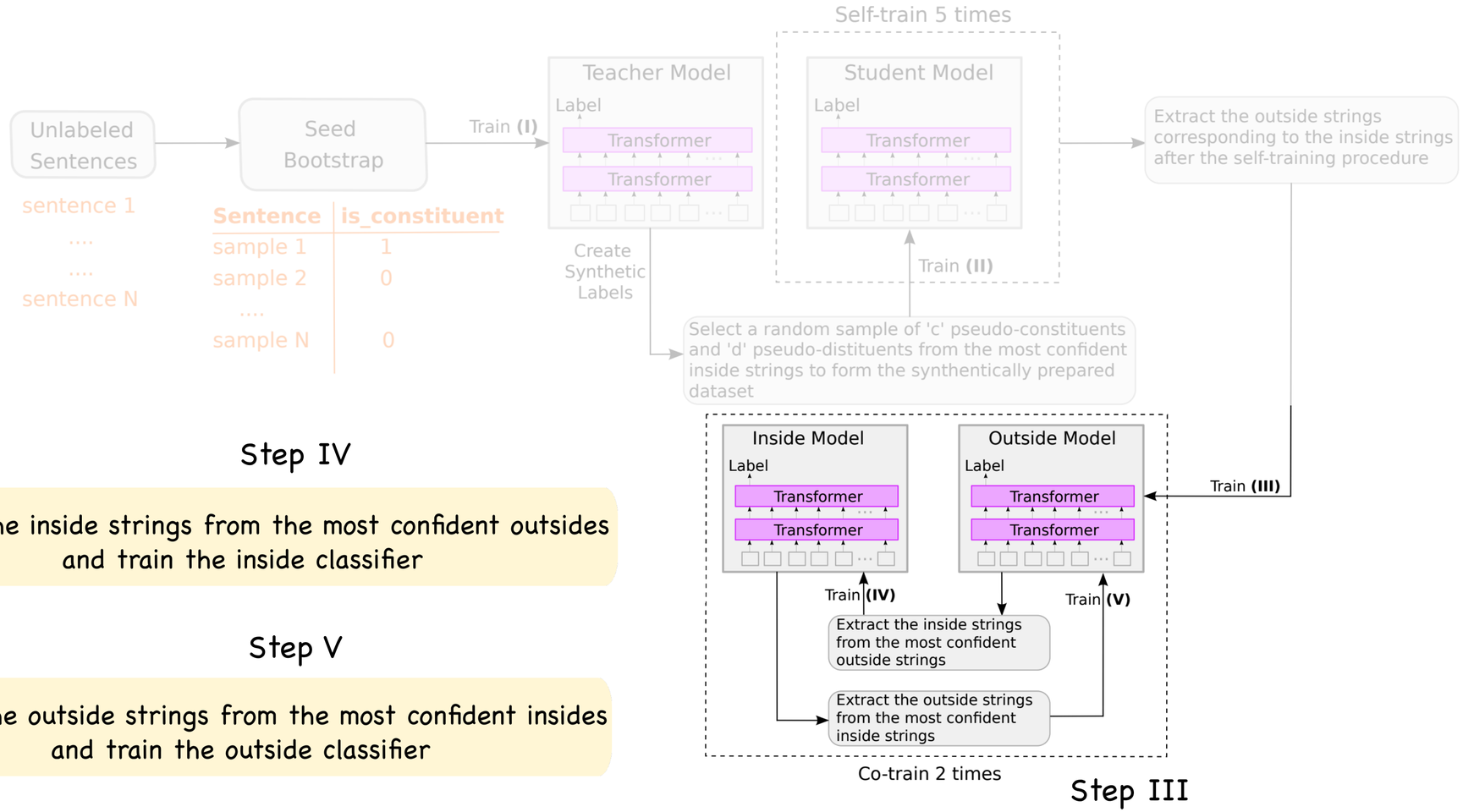
Implementation: Inside String with Self-training



Step I — Fine-tune a RoBERTa (base) model (teacher) on a downstream task using a cross-entropy loss after seed bootstrapping

Step II — Synthetically annotate this data using the teacher model and select top 'K' samples corresponding to each class to form the final synthetic dataset; We fine-tune a RoBERTa (base) model (student) on this dataset using hard labels and retrieve the outside strings from the most confident insides

Implementation: Inside-Outside Strings with Co-training



Step IV

Retrieve the inside strings from the most confident outsides and train the inside classifier

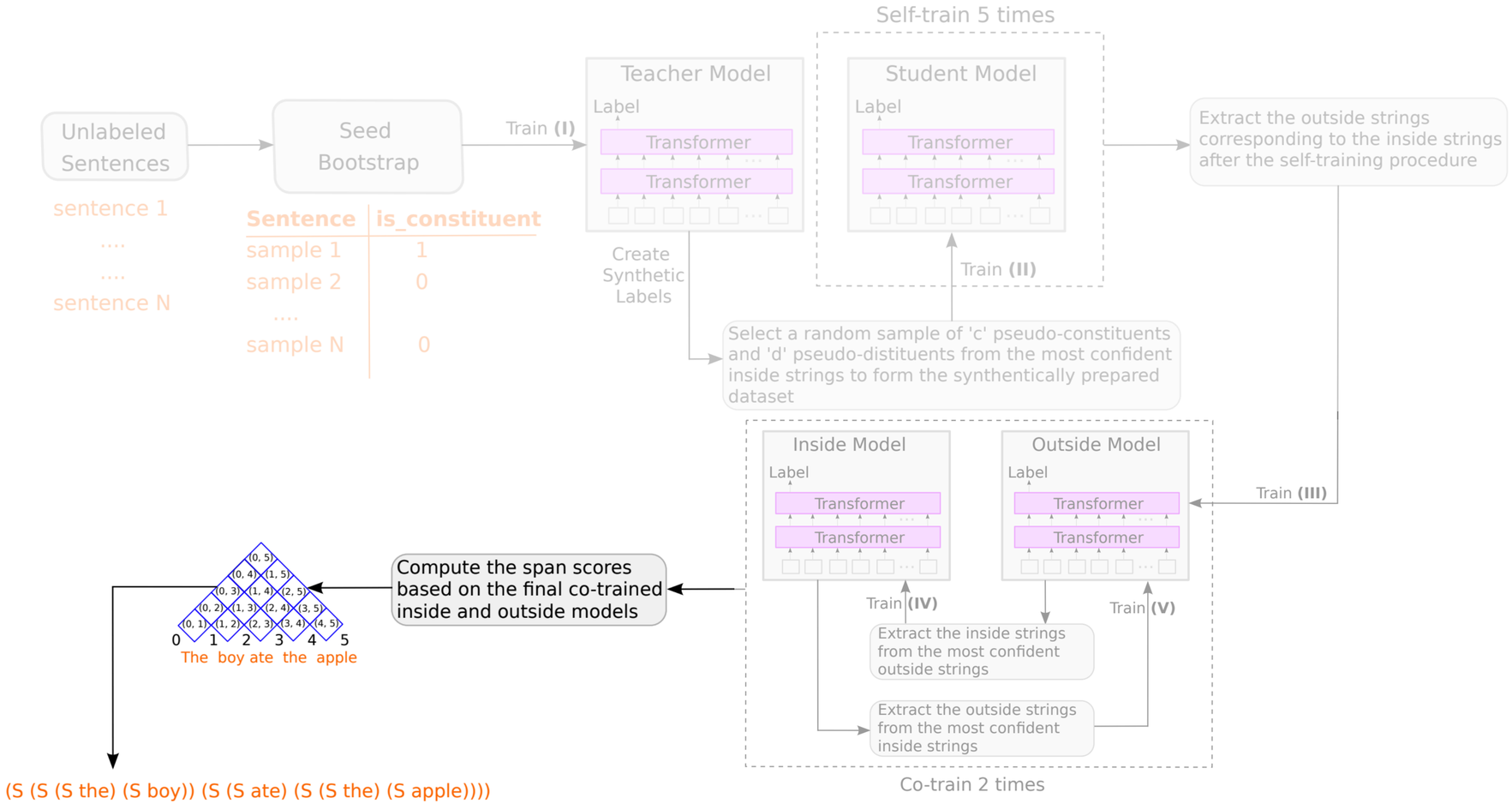
Step V

Retrieve the outside strings from the most confident insides and train the outside classifier

Step III

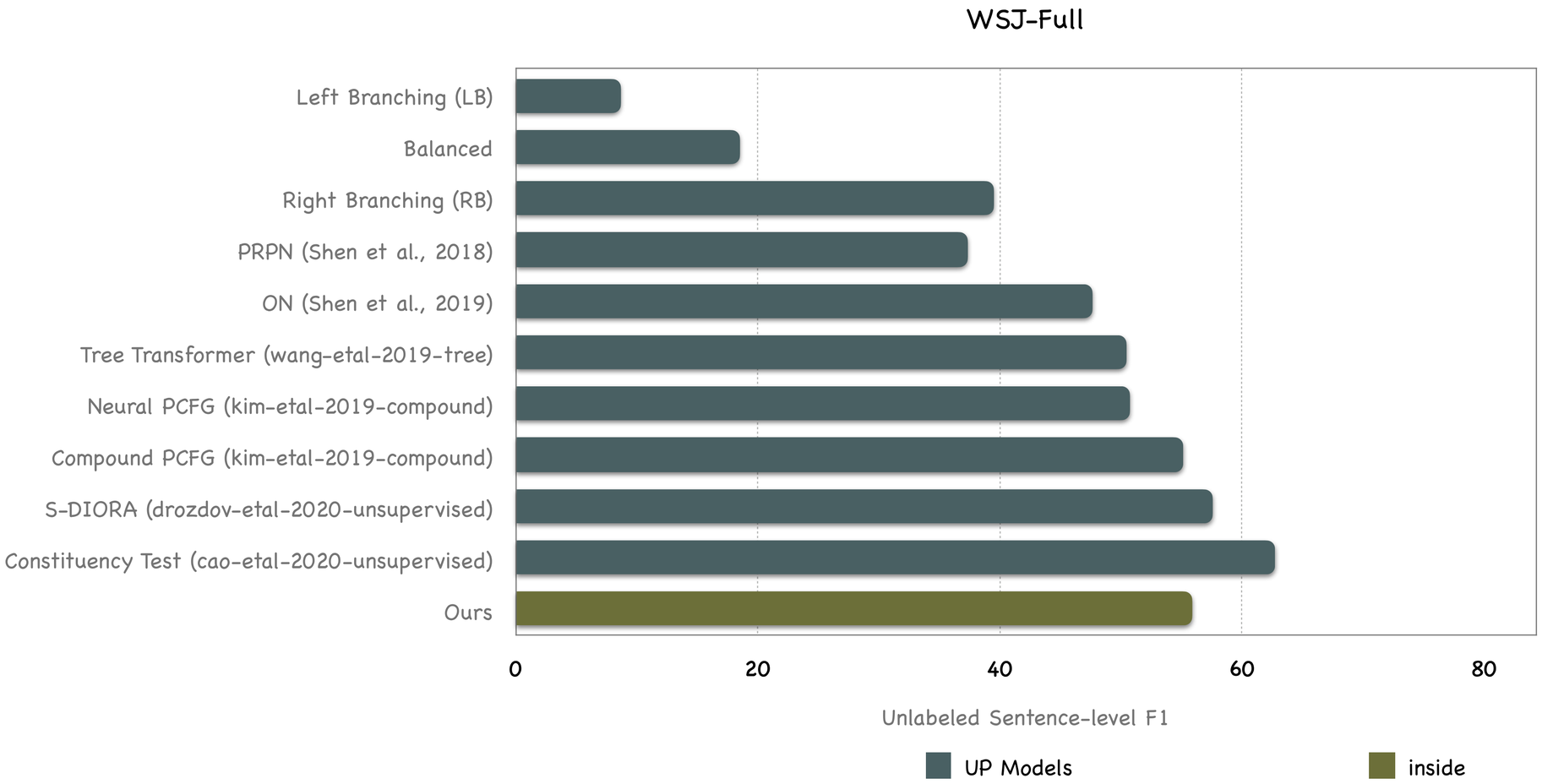
Train the outside classifier on these outside strings; We perform the co-training procedure for two iterations which follow a two-fold optimizing step

Implementation: Parsing Algorithm

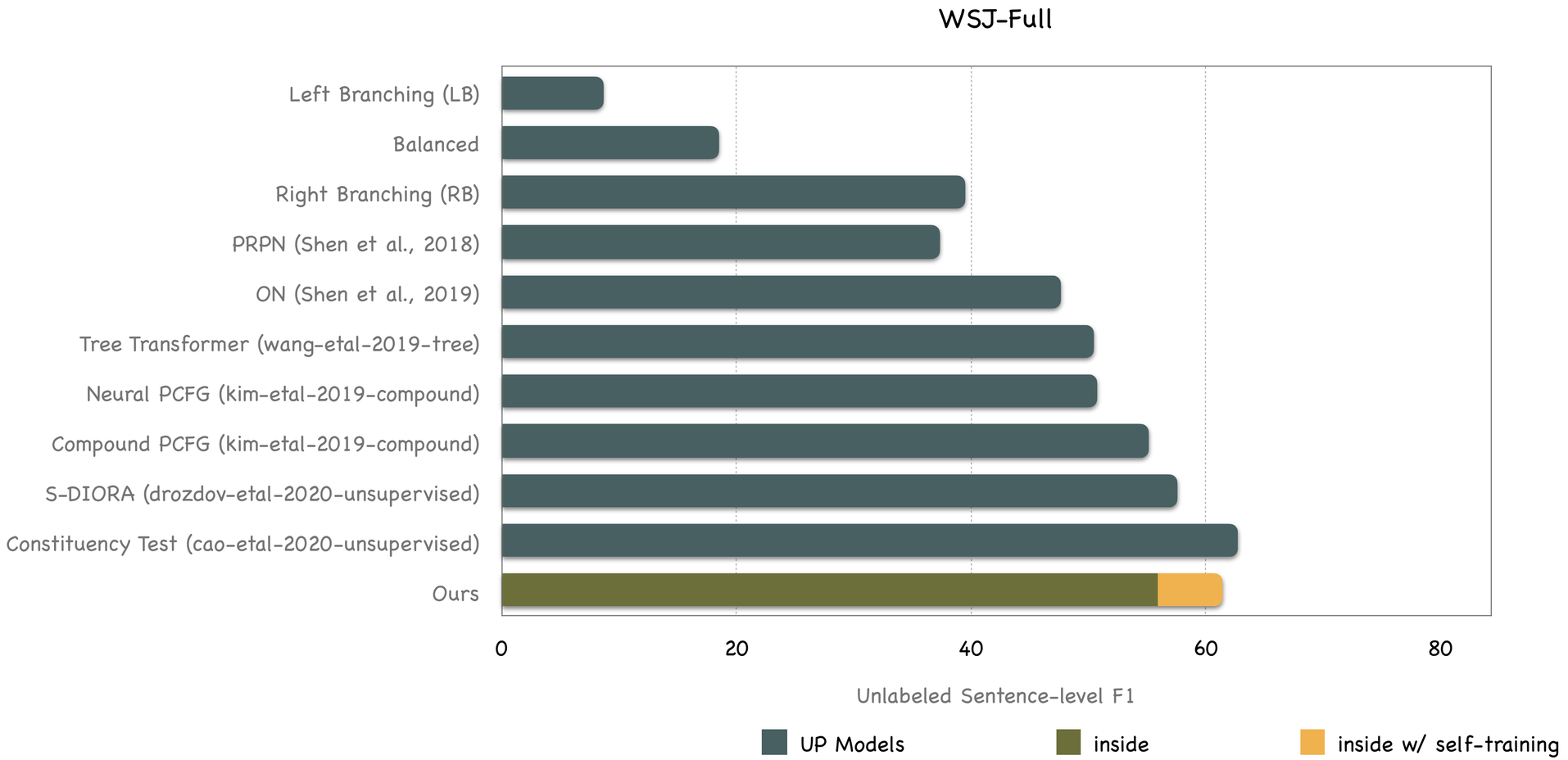


Viterbi form of the CYK algorithm to produce a globally optimized parse tree for each sentence

Results on the PTB test set

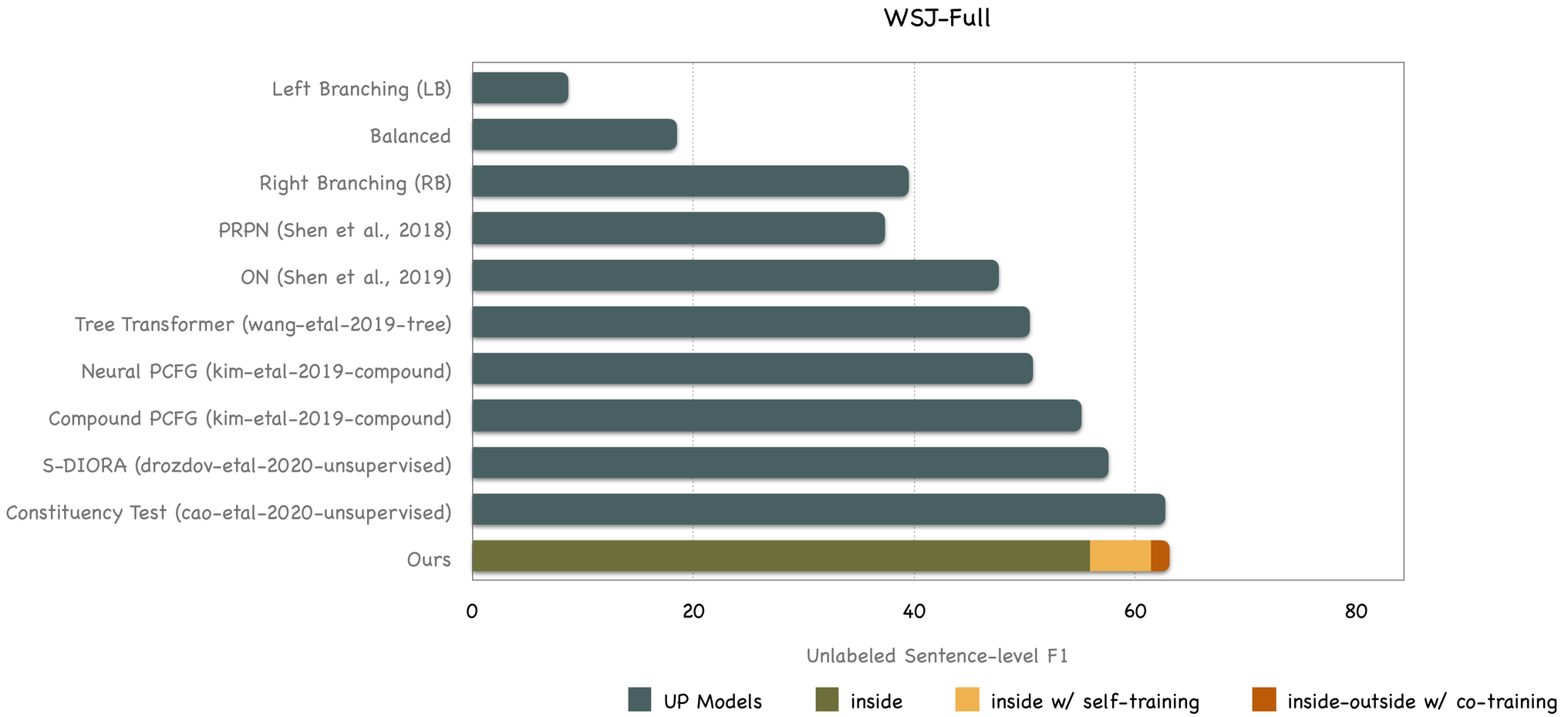


Results on the PTB test set



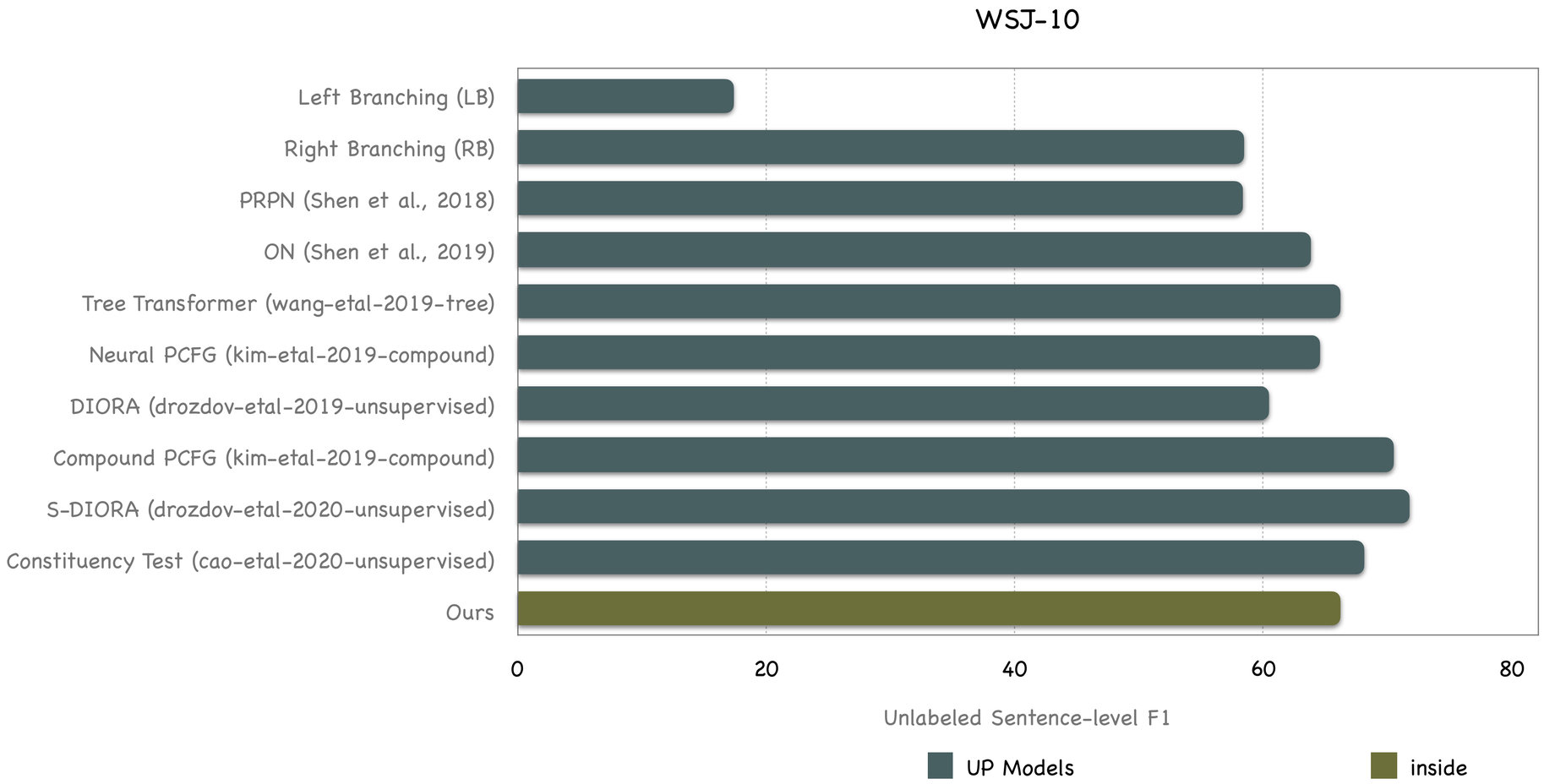
Inside model with self-training results in +5.5 F1 improvement compared to the vanilla (inside) model

Results on the PTB test set

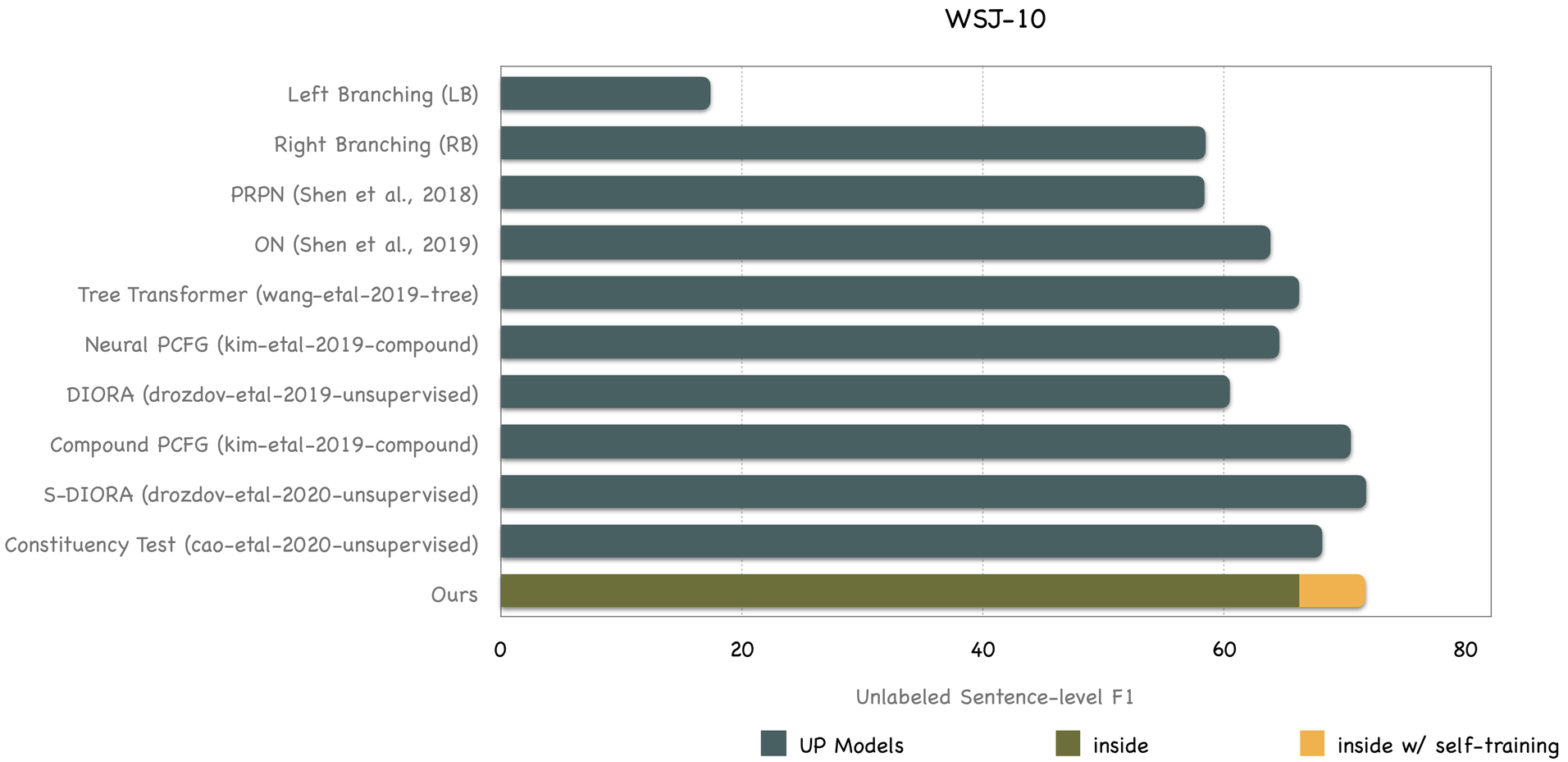


Inside-Outside model with co-training results in +7.2 F1 improvement compared to the vanilla (inside) model

Results on the PTB test set

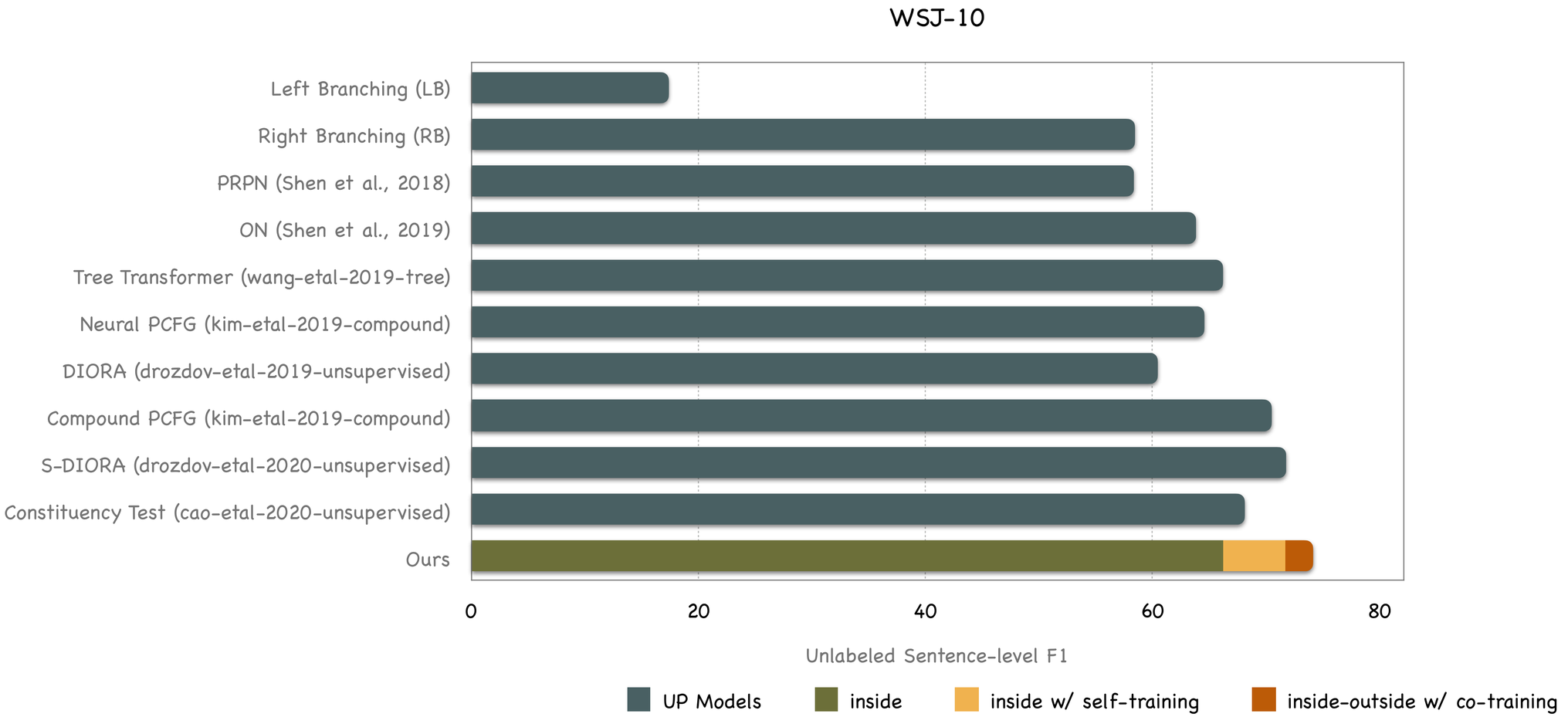


Results on the PTB test set



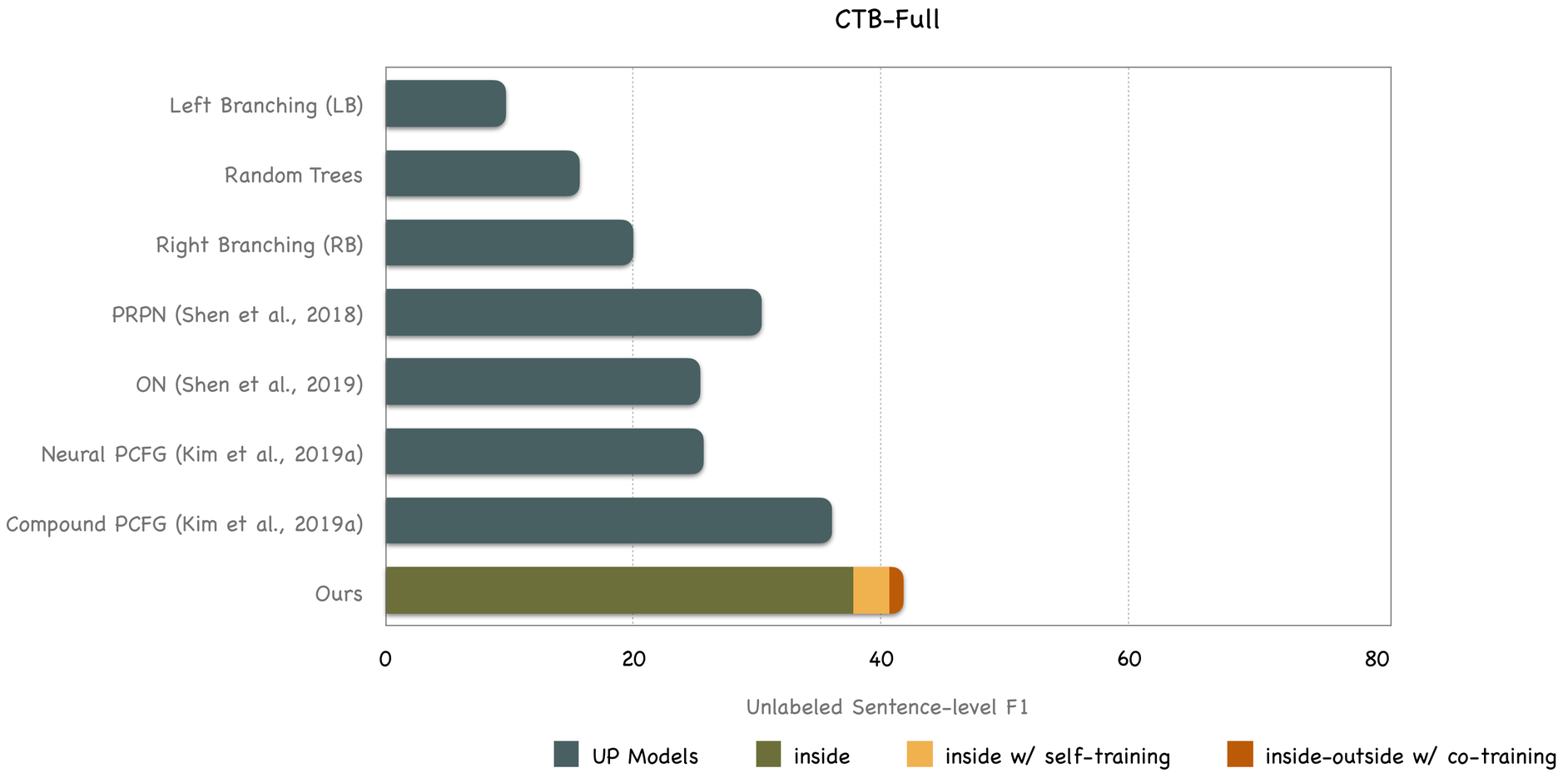
Inside model with self-training results in +5.5 F1 improvement compared to the vanilla (inside) model

Results on the PTB test set



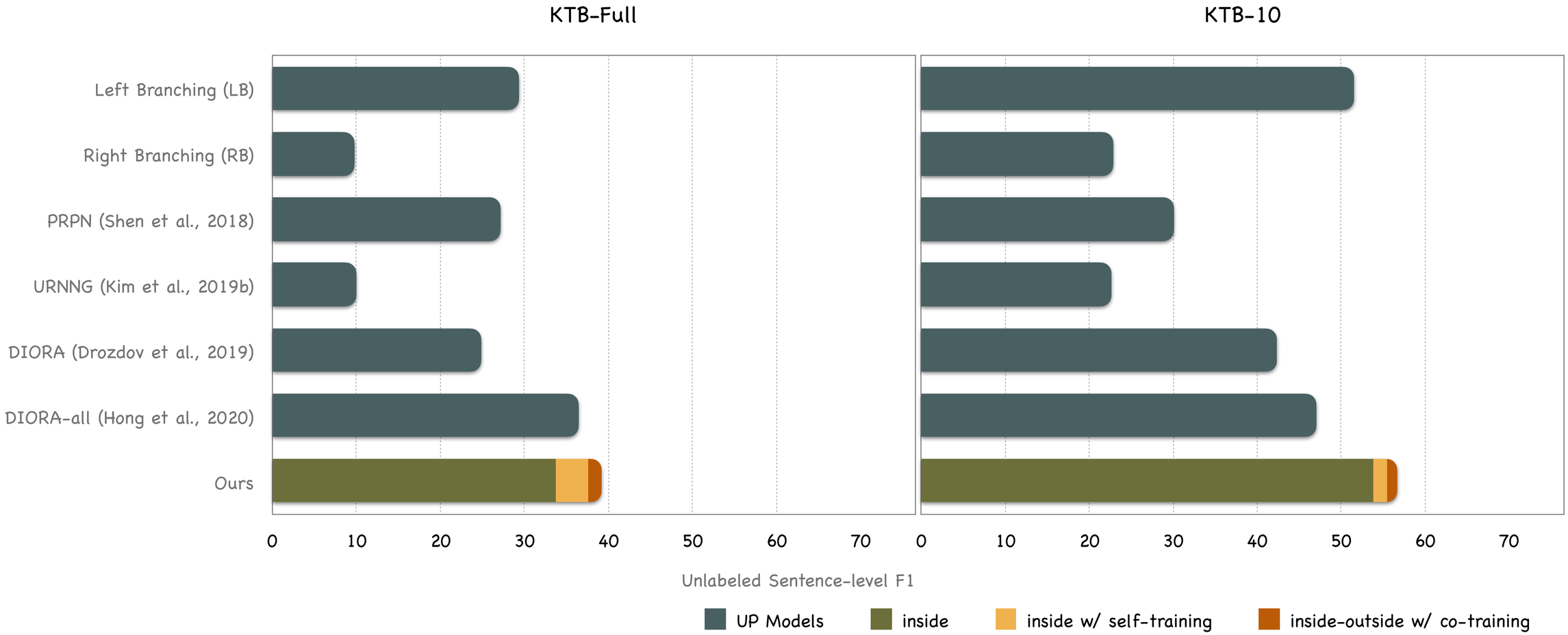
Inside-Outside model with co-training results in +8 F1 improvement compared to the vanilla (inside) model

Results on the CTB test set



Inside-Outside model with co-training results in +4 F1 improvement compared to the vanilla (inside) model

Results on the KTB test set



Inside-Outside model with co-training results in +5.5 F1 improvement compared to the vanilla (inside) model

Inside-Outside model with co-training results in +2.9 F1 improvement compared to the vanilla (inside) model

Conclusions

- Our parser has the ability to generalize to languages of known branching direction (left/right) and achieves new state-of-the-art results on three treebanks comprising both right- and left-branching languages.
- The use of inside and outside strings (inspired by the notion of inside and outside trees for the spectral learning of latent-variable PCFGs) is a crucial component in our pipeline.
- Employing semi-supervised learning techniques, i.e., self-training and co-training, to model interactions between the inside and outside classifiers results in an overall improved parsing performance.

Resources

- Code: https://github.com/Nickil21/weakly-supervised-parsing
- Models: https://huggingface.co/nickil/weakly-supervised-parsing
- Demo: https://huggingface.co/spaces/nickil/weakly-supervised-parsing
- Contact: https://nickilmaveli.com