Co-training an Unsupervised Constituency Parser with Weak Supervision

Nickil Maveli
@nickilmaveli
n.maveli@sms.ed.ac.uk
Shay B. Cohen
scohen@inf.ed.ac.uk

Institute for Language, Cognition and Computation, University of Edinburgh

Goal

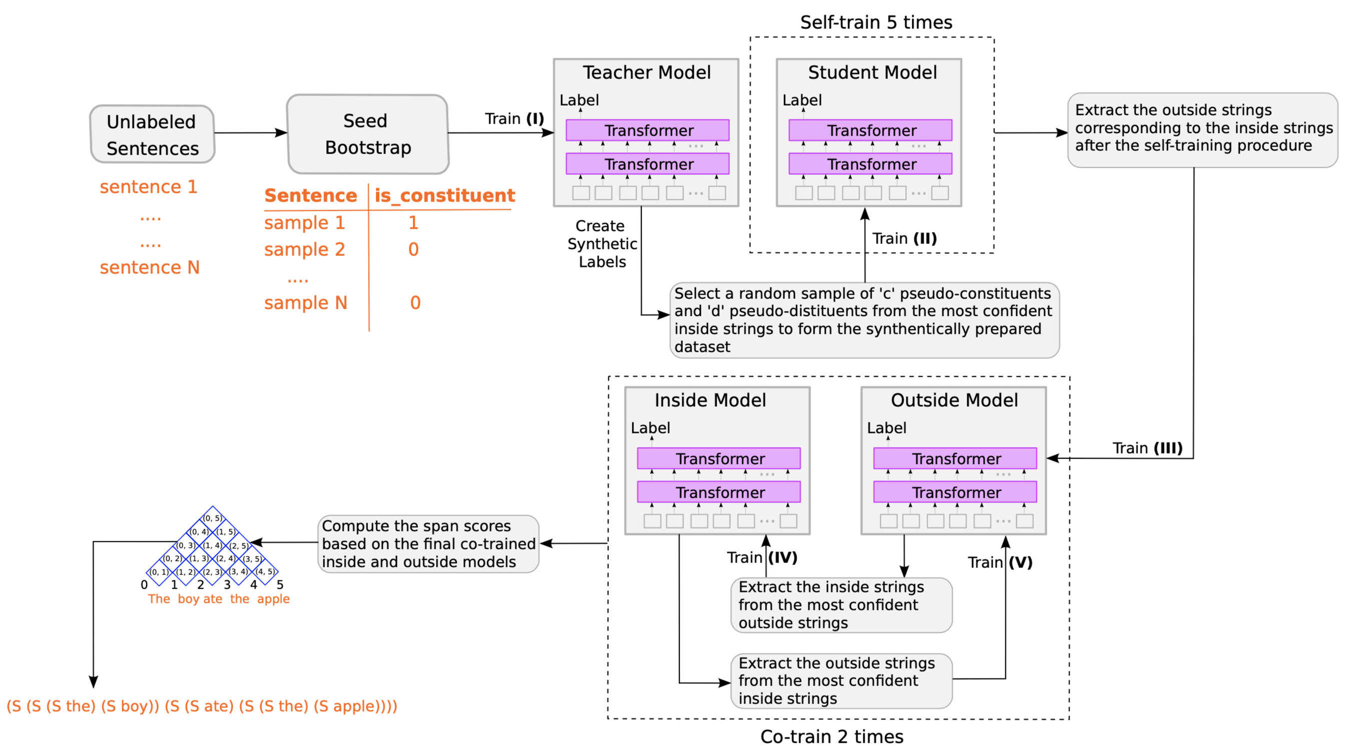
Induce parse trees from observed sentences alone without supervision.

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.

Proposed Approach

- We formulate the task of identifying constituents and distituents (referring to spans that are not constituents) in a sentence as a binary classification task by devising a `seed bootstrapping` strategy to convert the unlabeled data into a classification task.
- We build a sequence classification model by fine-tuning a Transformer-based PLM on the unlabeled training sentences to distinguish between the true and false `inside` strings of constituents.
- We use the highly-confident inside strings to produce the `outside` strings.
- Through the use of semi-supervised learning techniques, i.e., `self-training` and `co-training`, we jointly use both the inside and outside passes to enrich the model's ability to determine the breakpoints in a sentence.

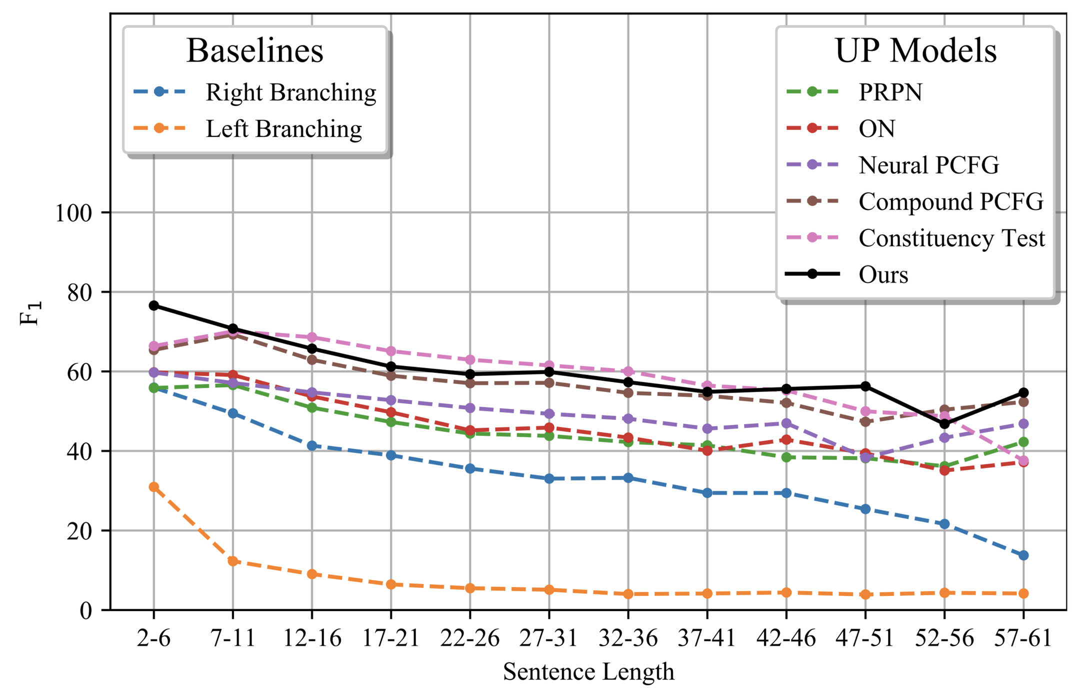


We perform the self-training procedure for five iterations which follow multiple steps:

1. Fine-tune a RoBERTa (base) model (teacher) on a downstream task using a cross-entropy loss after seed bootstrapping.
2. 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.
3. Train the outside classifier on these outside strings; We perform the co-training procedure for two iterations which follow a two-fold optimizing step.
4. Retrieve the inside strings from the most confident outsides and train the inside classifier.
5. Retrieve the outside strings from the most confident insides and train the outside classifier.

We use **semi-supervised** and **weakly-supervised** learning techniques to induce latent trees and achieve state-of-the-art results on three treebanks.

F1 vs. Sentence Length Plot



PTB Results

Model	WSJ-Full Mean	WSJ-Full Max	WSJ-10 Mean	WSJ-10 Max
Trivial Baselines:				
Left Branching (LB)	8.7		17.4	
Balanced	18.5			
Right Branching (RB)	39.5		58.5	
Unsupervised Parsing approaches:				
PRPN† (Shen et al., 2018)	37.4	38.1	58.4	–
URNNG* (Kim et al., 2019b)	–	45.4	–	–
ON† (Shen et al., 2019)	47.7	49.4	63.9	–
Tree Transformer†* (Wang et al., 2019)	50.5	52.0	66.2	–
Neural PCFG† (Kim et al., 2019a)	50.8	52.6	64.6	–
DIORA* (Drozdov et al., 2019)	–	58.9	60.5	–
Compound PCFG† (Kim et al., 2019a)	55.2	60.1	70.5	–
S-DIORA†* (Drozdov et al., 2020)	57.6	64.0	71.8	–
Constituency Test* (Cao et al., 2020)	62.8	65.9	68.1	–
Ours* (using inside)	55.9	57.2	66.2	–
Ours* (using inside w/ self-training)	61.4	64.2	71.7	–
Ours* (using inside and outside w/ co-training)	**63.1**	66.8	**74.2**	–
Oracle Binary Trees	84.3		82.1	

CTB Results

Model	CTB Mean	CTB Max
Trivial Baselines:		
Left Branching (LB)	9.7	
Random Trees	15.7	16.0
Right Branching (RB)	20.0	
Unsupervised Parsing approaches:		
PRPN (Shen et al., 2018)	30.4	31.5
ON (Shen et al., 2019)	25.4	25.7
Neural PCFG (Kim et al., 2019a)	25.7	29.5
Compound PCFG (Kim et al., 2019a)	36.0	39.8
Ours (using inside)	37.8	38.4
Ours (using inside w/ self-training)	40.6	41.7
Ours (using inside and outside w/ co-training)	**41.8**	43.3
Oracle Binary Trees	81.1	

KTB Results

Model	KTB-40 Mean	KTB-40 Max	KTB-10 Mean	KTB-10 Max
Trivial Baselines:				
Left Branching (LB)	29.4		51.6	
Right Branching (RB)	9.8		22.9	
Unsupervised Parsing approaches:				
PRPN (Shen et al., 2018)	27.2	31.8	30.1	33.6
URNNG (Kim et al., 2019b)	10	10.2	22.7	22.7
DIORA (Drozdov et al., 2019)	24.9	26.0	42.3	43.3
DIORA-all (Hong et al., 2020)	36.4	40.0	47.1	48.9
Ours (using inside)	33.7	36.3	53.8	55.9
Ours (using inside w/ self-training)	37.6	39.8	55.5	58.2
Ours (using inside and outside w/ co-training)	**39.2**	41.1	**56.7**	59.1
Upper Bound	76.5		76.6	

Key Findings

- 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 to model interactions between the inside and outside classifiers results in an overall improved parsing performance.

informatics
Edinburgh NLP