A Regularization-Based Framework for Bilingual Grammar Induction

Yong Jiang[◇] · Wenjuan Han[†] · Kewei Tu[†]

[†] School of Information Science and Technology, ShanghaiTech University
[◇] Alibaba DAMO Academy, Alibaba Group
[◇] work was done when at ShanghaiTech University

Motivation

We observe that learning the unsupervised Convex-MST model on the English corpus and then directly applying it to parse other languages produces surprisingly good results.

Proposed Approach

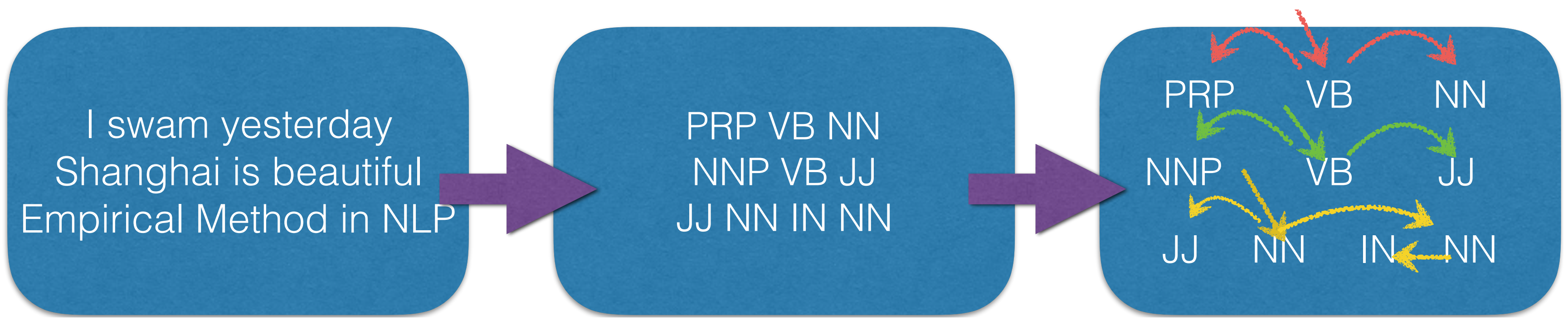
We propose three regularization methods for bilingual grammar induction that encourage similarity between models of two languages in terms of model parameters, dependency edge scores, and parse trees respectively.

Empirical Results

Empirical results on multiple languages show that our methods outperform strong baselines.

Unsupervised Dependency Parsing

Dependency parsing pipeline (POS tagging step can be obmitted):



Unsupervised dependency parsing:

- No access to golden trees.
- Usually taking POS tags as inputs.
- Typically incorporating different types of inductive biases.

Observations: Experiments of Direct Transfer Approach

Transferring the Convex-MST model trained on the English dataset to other languages.

Language	German	English	Spanish	French	Indonesian
Code	DE	EN	ES	FR	ID
C-MST	**60.2**	62.3	**68.8**	**72.3**	**69.7**
D-Tran	59.9	–	65.3	67.8	45.7
Δ	-0.3	–	-3.5	-4.5	-24.0
Language	Italian	Japanese	Korean	Portuguese	Swedish
Code	IT	JA	KO	PTBR	SV
C-MST	64.3	**57.5**	**59.0**	**68.3**	66.2
D-Tran	63.1	54.6	50.0	66.2	**67.8**
Δ	-1.2	-2.9	-9.0	-2.1	+1.6

C-MST: Convex-MST model (Grave and Elhadad, 2015), **D-Tran**: Direct Transfer

Findings:

- The dependency accuracy of **D-Tran** on each language is often very close to the accuracy of the model specifically trained on the corpus of that language
- For Swedish, the accuracy of direct transfer is even better than that of the specifically trained model.

Framework for Unsupervised Dependency Parsing

Regardless of model architectures, current unsupervised dependency models usually use the following form of objective function,

$$J(\mathbf{w};\mathcal{X}) = \sum_{\mathbf{x}\in\mathcal{X}} \bigcirc_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})}\Big(D(\mathbf{w},\mathbf{x},\mathbf{y}) + R(\mathbf{w})\Big)$$

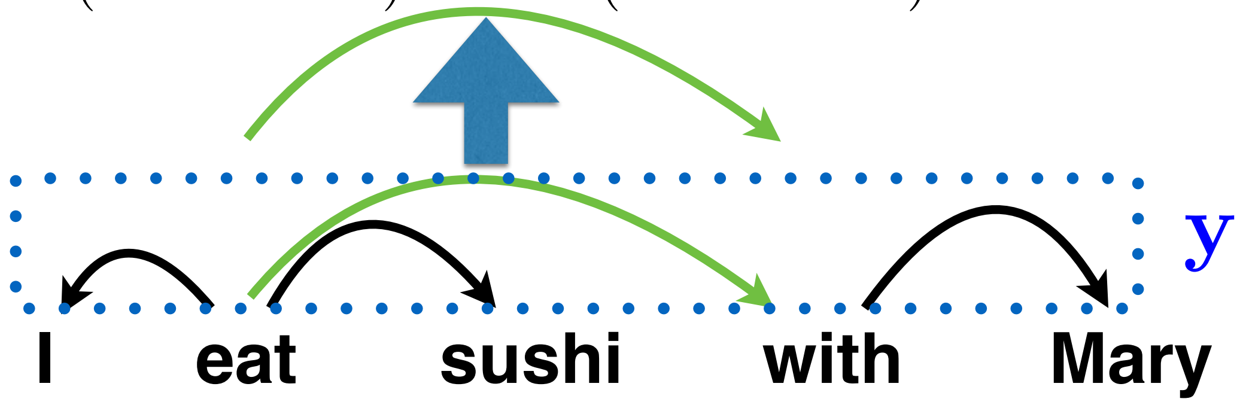
- $\mathcal{Y}$ is the set of all possible dependency tree
- $\mathbf{w}$ is the model parameter
- $\mathcal{X}$ is the unlabeled training corpus
- D is the measurement between the parse $\mathbf{y}$ and model prediction on sentence $\mathbf{x}$
- $R(\mathbf{w})$ is the regularization term of parameter $\mathbf{w}$
- $\bigcirc \in \{\mathbf{min}, \sum\}$ is an operator.

The Specifications of $\bigcirc$, D and R for several widely used models:

Parsers	$\bigcirc$	D	R
DMV	$\sum$	negative log likelihood	-
Convex-MST	**min**	ℓ_2 distance	ℓ_2 norm
LC-DMV	$\sum$	negative log likelihood	ℓ_2 norm
NDMV	$\sum$, **min**	negative log likelihood	-
CRFAE	**min**	negative conditional log likelihood	ℓ_1 norm
D-NDMV	$\sum$, **min**	negative (conditional) log likelihood	-

Bilingual Knowledge Sharing

$$\mathbf{s}(\text{eat} \rightarrow \text{with}) = \mathbf{w}^T\mathbf{f}(\text{eat} \rightarrow \text{with})$$



Three levels of knowledge:

- the parameter $\mathbf{w}$,
- the edge score $\mathbf{s}$,
- the parse $\mathbf{y}$.

Vanilla Joint Training:

$$J(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s,\mathcal{X}_t) = J(\mathbf{w}_s;\mathcal{X}_s) + J(\mathbf{w}_t;\mathcal{X}_t)$$

Regularization of Weight Parameters (W-Reg)

$$J(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s,\mathcal{X}_t) = J(\mathbf{w}_s;\mathcal{X}_s) + J(\mathbf{w}_t;\mathcal{X}_t) + \lambda||\mathbf{w}_s - \mathbf{w}_t||_2^2$$

Regularization on Edge Scores (E-Reg)

$$J(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s,\mathcal{X}_t) = J(\mathbf{w}_s;\mathcal{X}_s) + J(\mathbf{w}_t;\mathcal{X}_t) + \lambda\sum_{\mathbf{x}\in\mathcal{X}'}\sum_{(h,m)\in\mathcal{G}(\mathbf{x})}||\mathbf{s}_{\mathbf{w}_s}(\mathbf{x},h,m) - \mathbf{s}_{\mathbf{w}_t}(\mathbf{x},h,m)||_2^2$$

where $\mathcal{X}' = \mathcal{X}_s \cup \mathcal{X}_t$. $\mathcal{G}(\mathbf{x})$ is the dependency graph of sentence $\mathbf{x}$.

Regularization on Parse Trees (T-Reg)

$$J'(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s) = \sum_{\mathbf{x}\in\mathcal{X}_s} \bigcirc_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})}\Big(D(\mathbf{w}_s,\mathbf{x},\mathbf{y}) + \lambda \underbrace{D(\mathbf{w}_t,\mathbf{x},\mathbf{y})}_{\text{T-Reg term}} + R(\mathbf{w}_s)\Big)$$

$$J'(\mathbf{w}_t,\mathbf{w}_s;\mathcal{X}_t) = \sum_{\mathbf{x}\in\mathcal{X}_t} \bigcirc_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})}\Big(D(\mathbf{w}_t,\mathbf{x},\mathbf{y}) + \lambda \underbrace{D(\mathbf{w}_s,\mathbf{x},\mathbf{y})}_{\text{T-Reg term}} + R(\mathbf{w}_t)\Big)$$

$$J(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s,\mathcal{X}_t) = J'(\mathbf{w}_s,\mathbf{w}_t;\mathcal{X}_s) + J'(\mathbf{w}_t,\mathbf{w}_s;\mathcal{X}_t)$$

Experiments

Dataset: Universal Treebanks Version 2.

- Training data: Sentence length no more than 10.
- Testing data: Sentence length no more than 10 and all sentences.

CODE	C-MST	D-TRAN	W-REG	E-REG	T-REG
DE	60.2	-0.3	-0.2	**+0.2**	-0.2
ES	68.8	-3.5	-3.5	**+0.8**	+0.3
FR	72.3	-4.5	-3.8	**+0.3**	**+0.3**
ID	**69.7**	-24.0	-21.4	-0.6	-1.2
IT	64.3	-1.2	-0.4	+0.3	**+1.2**
JA	57.5	-2.9	-3.4	+0.9	**+2.3**
KO	59.0	-9.0	-9.5	+1.3	**+1.9**
PTBR	68.3	-2.1	-2.1	+0.2	**+0.3**
SV	66.2	+1.6	+1.6	**+2.7**	+2.6
Avg	65.14	-5.10	-4.74	+0.68	**+0.83**
Avg-All	56.16	-4.63	-4.29	+0.47	**+0.56**

Figure 1. Transfer grammar induction from English to the other language

- **Avg:** Averaged directed dependency accuracy of sentences with length $\leq$ 10.
- **Avg-All:** Averaged directed dependency accuracy of all sentences.
- **COMB:** Model trained by concating data from two languages.

PAIR	CODE	BASE	COMB	W-REG	E-REG	T-REG
EN-DE	EN	62.3	+0	**+0.5**	+0.1	+0.2
	DE	60.2	+0.1	-0.8	**+0.5**	+0.4
EN-ES	EN	62.3	+0.2	**+0.7**	+0.4	**+0.7**
	ES	68.8	-1.8	**+1.4**	+1.1	+0.8
EN-FR	EN	62.3	+0.4	**+0.5**	+0.1	+0.1
	FR	72.3	-3.6	-0.6	**+1.4**	**+1.4**
EN-ID	EN	62.3	+0.7	**+1.3**	+0.1	+0.7
	ID	69.7	-19.2	-2.1	+0.5	**+0.9**
EN-IT	EN	62.3	+0.7	+0.1	+0.6	**+1.0**
	IT	64.3	+0.7	**+1.8**	+0.4	+0.8
EN-JA	EN	62.3	+1.0	**+0.9**	-0.1	-0.2
	JA	57.5	-4.3	-0.2	**+2.7**	+2.4
EN-KO	EN	62.3	+0.3	**+1.4**	+0.5	+1.1
	KO	59.0	-3.8	+0.1	**+1.0**	+0.4
EN-PTBR	EN	62.3	+1.0	**+0.8**	+0.6	+0.7
	PT-BR	68.3	-0.9	**+1.4**	**+1.4**	+1.2
EN-SV	EN	62.3	-0.5	-0.9	**+0.1**	**+0.1**
	SV	66.2	+1.7	**+1.9**	+1.0	+1.1
Avg	EN	62.30	+0.42	**+0.58**	+0.23	+0.49
	Other	65.14	-3.46	+0.32	**+1.11**	+1.04
Avg-All	EN	52.10	-0.48	-0.40	+0.11	**+0.42**
	Other	56.16	-3.05	+0.64	**+1.55**	+1.21

Figure 2. Results of bilingual grammar induction on test sentences no longer than 10