Towards Semi-Supervised Learning for Deep Semantic Role Labeling

Sanket Vaibhav Mehta*, **Jay-Yoon Lee***, **Jaime Carbonell**

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

* denotes equal contribution

Carnegie Mellon University
School of Computer Science

Summary

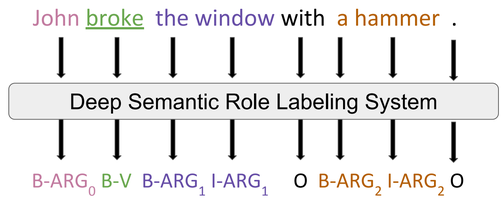
We propose to inject structural constraint on deep Semantic Role Labeling (SRL) models and present a novel Semi-Supervised Learning formulation. Our proposed method leverages syntactic parse trees by introducing them as hard constraints on SRL output during training. The proposed method significantly improves both constraint satisfaction and overall F1, especially on low-resource settings.

Setting

Formulate SRL as a BIO tagging problem

Input: a sentence-predicate pair

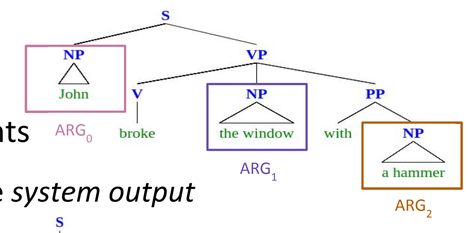
Output: a sequence of tags



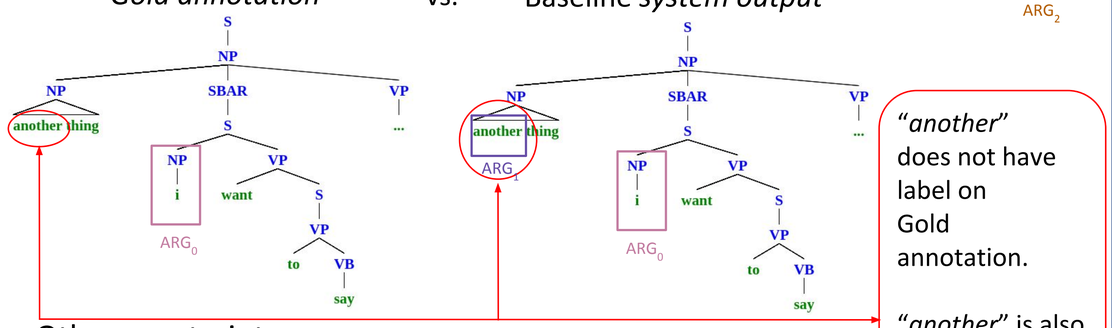
Motivation

Structural Constraints

- Syntactic constraints :
  - SRL argument spans ⊂ parse constituents



Gold annotation vs. Baseline *system output*



"*another*" does not have label on Gold annotation.

"*another*" is also not a valid parse tree constituent.

- Other constraints
  - BIO constraints : B-ARG$_0$ I-ARG$_1$ (Viterbi decoding handles)
  - SRL constraints :
    - Unique core (U) / Continuation (C) / Reference (R) roles

Baseline Method

Training

- Locally normalized models
- Cross-entropy objective (NLL)

$$\mathcal{L}(\mathbf{w}) = -\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}} \sum_{i=1}^{|\mathbf{y}|} \log p(y_i|\mathbf{x};\mathbf{w}).$$

Inference

- Enforcing structural constraints during inference by augmenting penalty terms for constraint violations c(x,y).

$$\mathbf{f}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}) - \sum_{c\in C} c(\mathbf{x},\mathbf{y}_{1:i}) \qquad \hat{\mathbf{y}} = \arg\max_{\mathbf{y}'\in\Omega^n} \mathbf{f}(\mathbf{x},\mathbf{y}')$$

Proposed Method

Enforcing syntactic consistency (during training)

- Disagreement rate:

$$d(\mathbf{x},\mathbf{y}) = \frac{|\text{disagreeing-spans}(\mathbf{x},\mathbf{y})|}{|\text{srl-spans}(\mathbf{y})|}$$

- Syntactic Inconsistency (SI) score: $s(\mathbf{x},\mathbf{y}) = 2 \times d(\mathbf{x},\mathbf{y}) - 1$
- Objectives
  - Syntactic Inconsistency objective:

$$= s(\mathbf{x},\hat{\mathbf{y}}^{(t)}) \sum_{i=1}^{|\hat{\mathbf{y}}^{(t)}|} \log p(\hat{y}_i^{(t)}|\mathbf{x};\mathbf{w}^{(t)}) \quad \text{s.t. } \hat{\mathbf{y}} = \arg\max_{\mathbf{y}'\in\Omega^n} \mathbf{f}(\mathbf{x},\mathbf{y}')$$

  - Joint objective:
  - Semi-Supervised Joint objective :

SRL-***labeled*** data $(\mathbf{x}_m = \mathbf{x}_k)$

$$-\alpha_1 \sum_{i=1}^{|\mathbf{y}|} \log p\,(y_i \mid \mathbf{x}_k;\mathbf{w}) + \alpha_2 s(\mathbf{x},\hat{\mathbf{y}}^{(t)}) \sum_{i=1}^{|\hat{\mathbf{y}}^{(t)}|} \log p(\hat{y}_i^{(t)} \mid \mathbf{x}_m;\mathbf{w}^{(t)})$$

SRL-***labeled*** data — SRL-***unlabeled*** data $(\mathbf{x}_m \neq \mathbf{x}_k)$

Questions

Q1. How well does the ***baseline*** model produce syntactically ***consistent*** outputs?
Q2. Does our ***proposed method*** actually ***enforce*** syntactic constraints?
Q3. Does our proposed method ***enforce*** syntactic constraints ***without compromising*** the quality of the system?
Q4. How well does our ***SSL formulation perform***, especially in ***low-resource*** scenarios?
Q5. What is the difference in enforcing the syntactic constraints in the ***training time vs. decoding time*** ?

Experiment

Data

- OntoNotes v5.0, CoNLL-2012 shared task (278k train, 38.3k dev, 25.6k test)
- Gold predicates and gold parse constituents are given (~10% noisy annotations)

Models

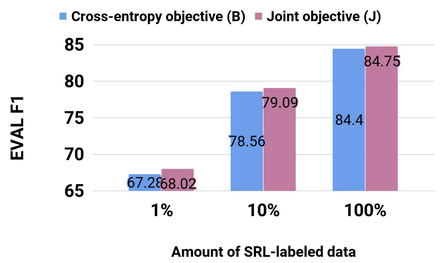
BX	model trained with X% of the SRL-labeled data
BX, JX	models trained with cross-entropy and joint objective respectively
BX-SIUx	model trained with SI loss with XxU amount of the SRL-unlabeled data used for further training pre-trained BX
BX-SIUx, BX-JUx	models trained with SI-loss and joint loss respectively

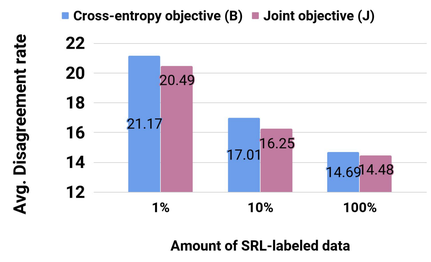
Evaluation

- EVAL F1 : Overall F1-score on CoNLL-2012-test set
- Average disagreement rate : Average d(x,y) across examples.

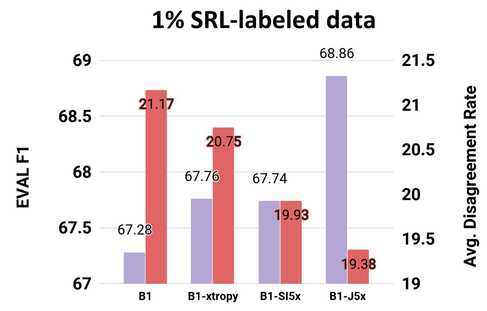
Experiment Result

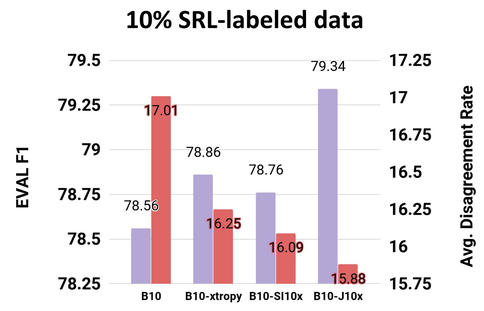
- Evidence for answering Q1, Q2 and Q3 favorably



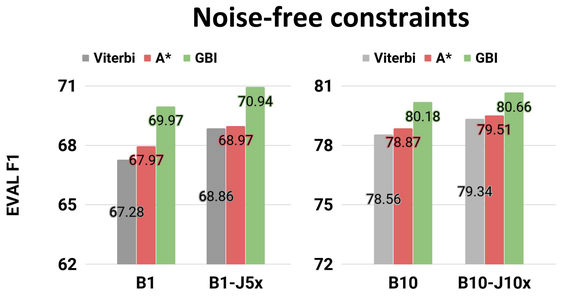


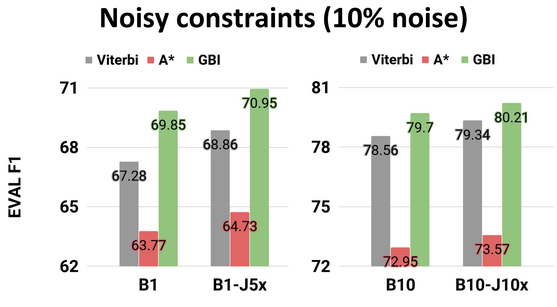
- Evidence for answering Q3 and Q4 favorably





- Evidence for answering Q5





Discussion

- Using system predicted parse constituents (tri-training) and predicates
- New experiments of SSL on full data made available.

Parse Tree & Predicate	B1-J5x		B10-J10x		J100-J3x	
	EVAL F1	Avg. Disagreement	EVAL F1	Avg. Disagreement	EVAL F1	Avg. Disagreement
Gold	68.86 (+1.58)	19.38 (-1.79)	79.34 (+0.78)	15.88 (-1.13)	-	-
System predicted	68.57 (+1.29)	19.73 (-1.44)	79.01 (+0.45)	16.39 (-0.62)	84.87 (+0.47)	14.23 (-0.46)