

Backpropagating through Structured Argmax using a SPIGOT

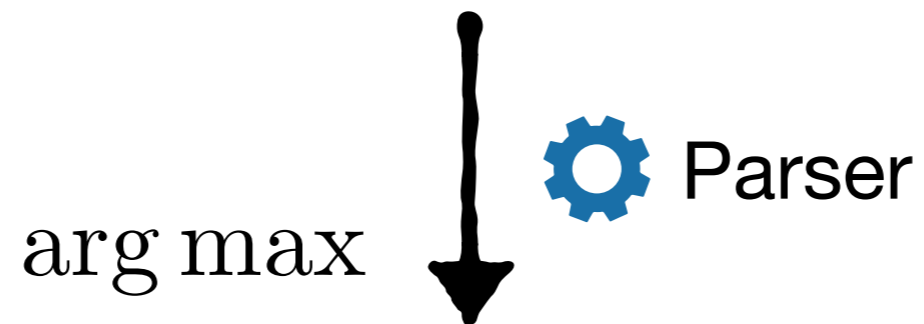
Hao Peng, Sam Thomson, Noah A. Smith



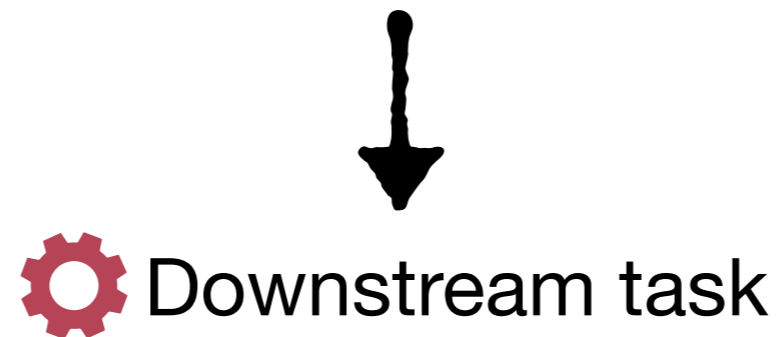
@ACL
July 17, 2018

Overview

Shareholders took their money



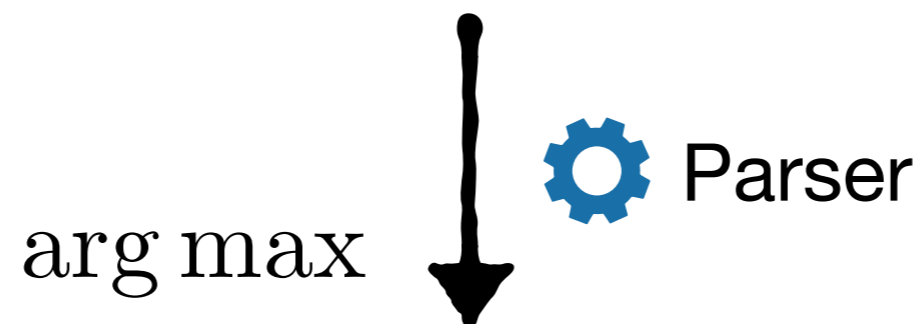
Shareholders took their money



Loss \mathcal{L}

Overview

Shareholders took their money



Shareholders took their money

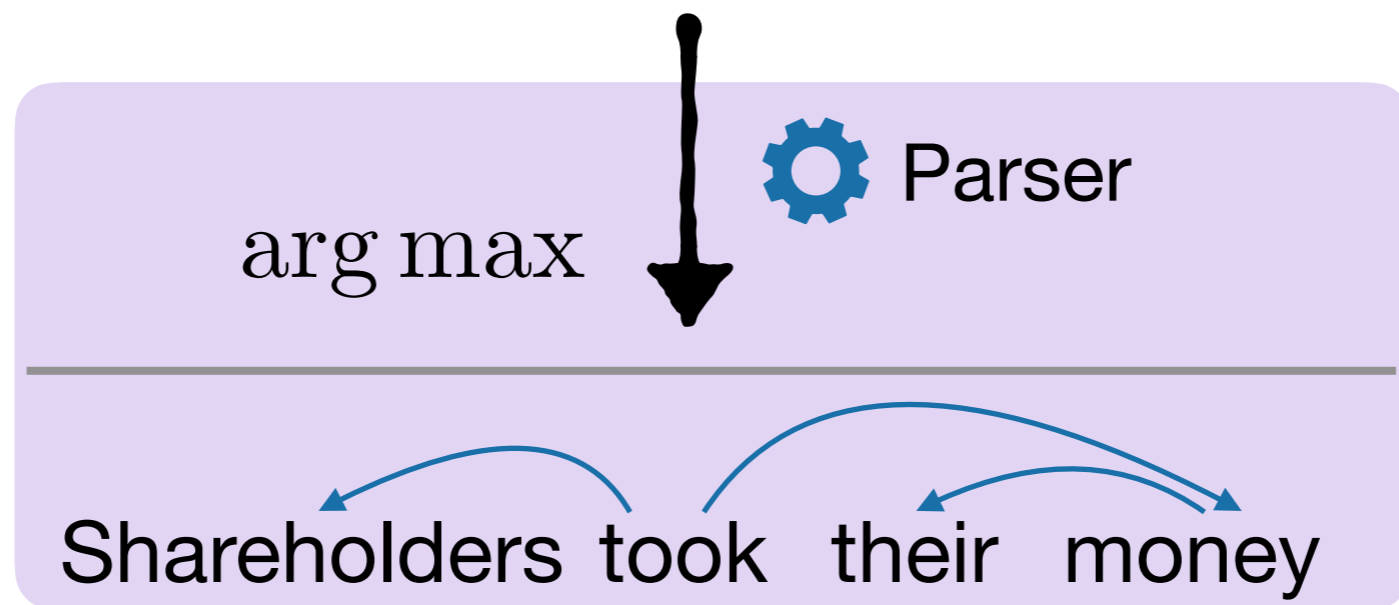


Loss \mathcal{L}

- Head token
Yang and Mitchell, 2017
- Tree-RNN
Tai et al., 2015
- Graph CNN
Kipf and Welling, 2017
- ...

Overview

Shareholders took their money

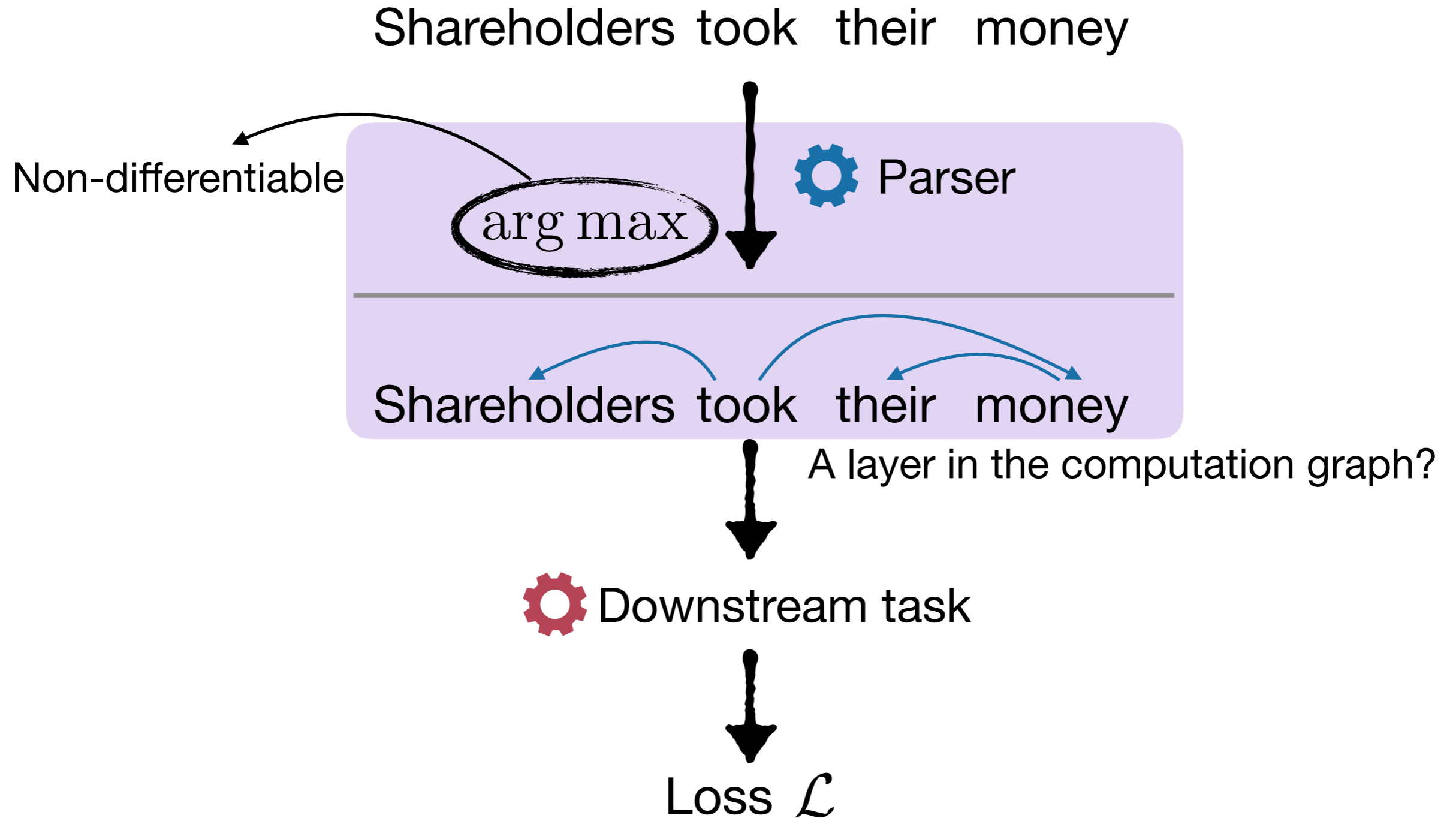


A layer in the computation graph?

 Downstream task

Loss \mathcal{L}

Overview



Overview

Aim

- Structured prediction as a layer.

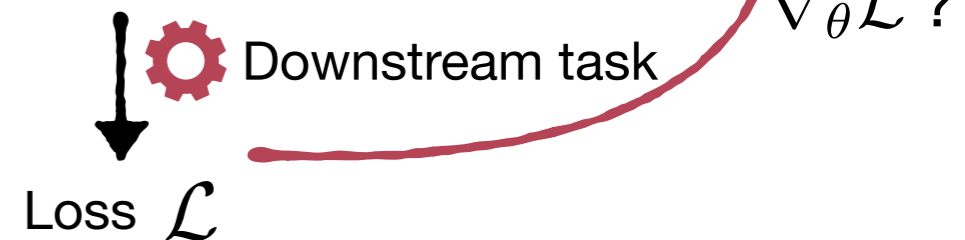
Motivation

- Structures help.
Ji and Smith, 2017; Oepen et al., 2017
- Linguistic structures may not be universally optimal.
Williams, 2017

Shareholders took their money



Shareholders took their money



Overview

Aim

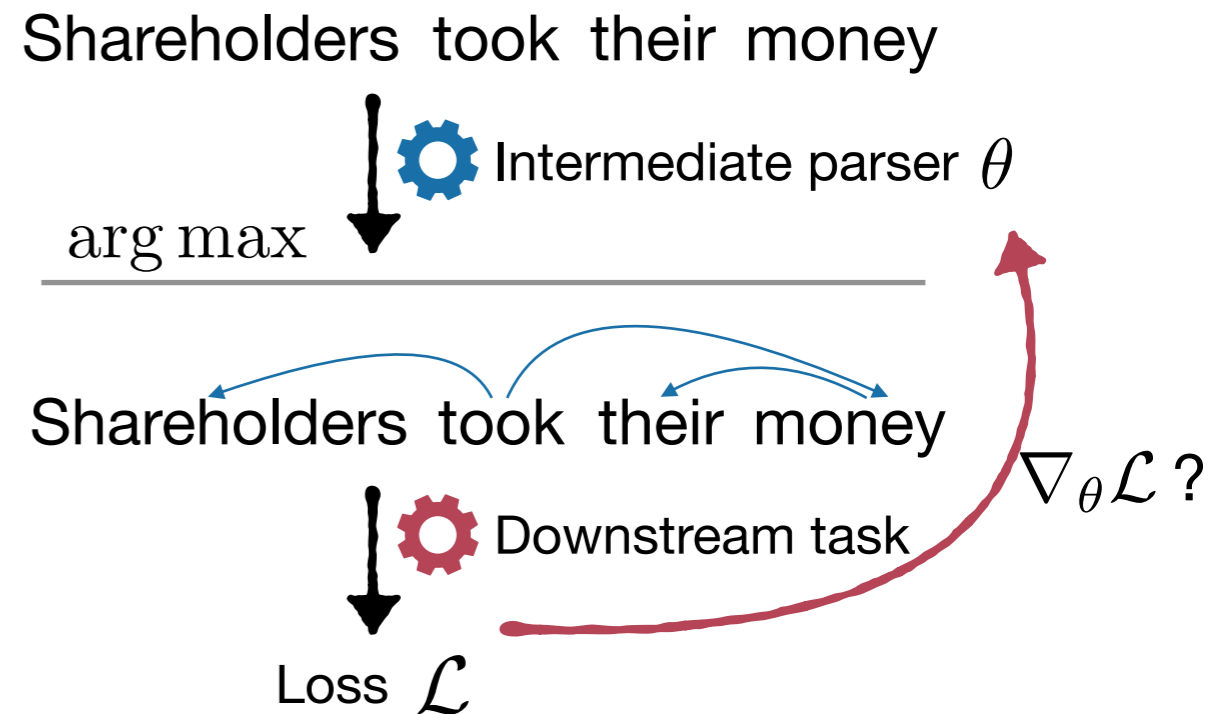
- Structured prediction as a layer.

Motivation

- Structures help.
Ji and Smith, 2017; Oepen et al., 2017
- Linguistic structures may not be universally optimal.
Williams, 2017

Challenges

- argmax is non-differentiable.



Overview

Aim

- Structured prediction as a layer.

Motivation

- Structures help.
Ji and Smith, 2017; Oepen et al., 2017
- Linguistic structures may not be universally optimal.
Williams, 2017

Challenges

- argmax is non-differentiable.

Shareholders took their money



arg max

Shareholders took their money



Loss \mathcal{L}

$\nabla_{\theta} \mathcal{L} ?$

A proxy

Method

Structured Prediction Intermediate
Gradients Optimization Technique

SPIGOT



Outline

- ❖ **Background: structured prediction as linear programs**
- ❖ Method: SPIGOT algorithm
- ❖ Experiments

Structured Prediction Reviewed

Input

Shareholders took their money

Output

Shareholders took their money

The diagram shows the same sentence "Shareholders took their money" as the input. Three blue curved arrows are drawn above the words to indicate segmentation. The first arrow starts above "Shareholders" and ends above "took". The second arrow starts above "took" and ends above "their". The third arrow starts above "their" and ends above "money".

Structured Prediction Reviewed


Input

Shareholders took their money

Score

$$S_{\theta} \left(\text{Shareholders took their money} \right)$$


||

$$\sum_{\text{arcs}} s_{\theta} \left(\text{head mod} \right)$$


Structured Prediction Reviewed

Input

Shareholders took their money

Score

$$\mathbf{s}_\theta = \left[s_\theta \left(\overset{\text{their}}{\curvearrowright} \text{money} \right), s_\theta \left(\overset{\text{took}}{\curvearrowright} \text{their} \right), s_\theta \left(\overset{\text{took}}{\curvearrowright} \text{money} \right), \dots, s_\theta \left(\overset{\text{their}}{\curvearrowright} \text{took} \right) \right]^\top$$

$$\mathbf{z} = \left[1?, 0?, 1?, \dots, 0? \right]^\top$$

Output

$$\arg \max \mathbf{z}^\top \mathbf{s}_\theta$$

s.t. \mathbf{Z} forms a tree

$\hat{\mathbf{z}}$

Shareholders took their money

Linear Programming Formulation

$\hat{\mathbf{z}}$ Shareholders took their money

$$\begin{array}{l} \text{arg max } \mathbf{z}^T \\ \text{s.t. } \mathbf{z} \text{ forms a tree} \end{array} \parallel \begin{bmatrix} s_\theta \text{ (their money)} \\ s_\theta \text{ (took their)} \\ s_\theta \text{ (took money)} \\ \vdots \\ s_\theta \text{ (their took)} \end{bmatrix}$$

$$\mathbf{Az} \leq \mathbf{b}$$

Linear Programming Formulation

$\hat{\mathbf{z}}$ Shareholders took their money

$\arg \max \mathbf{z}^T$

s.t. \mathbf{z} forms a tree

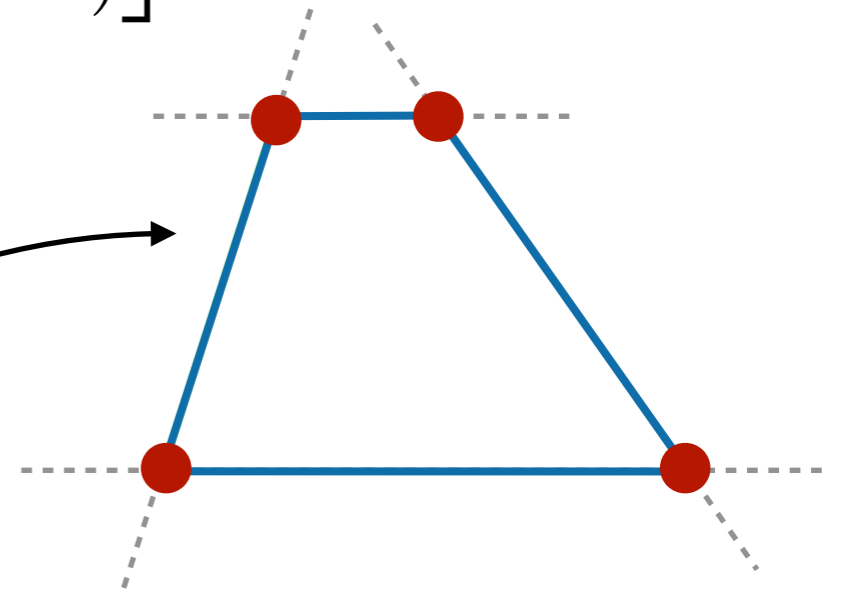
$z_i \in \{0, 1\}$

relaxation

$z_i \in [0, 1]$

$$\mathbf{Az} \leq \mathbf{b}$$

$$\begin{bmatrix} s_\theta \text{ (their money)} \\ s_\theta \text{ (took their)} \\ s_\theta \text{ (took money)} \\ \vdots \\ s_\theta \text{ (their took)} \end{bmatrix}$$

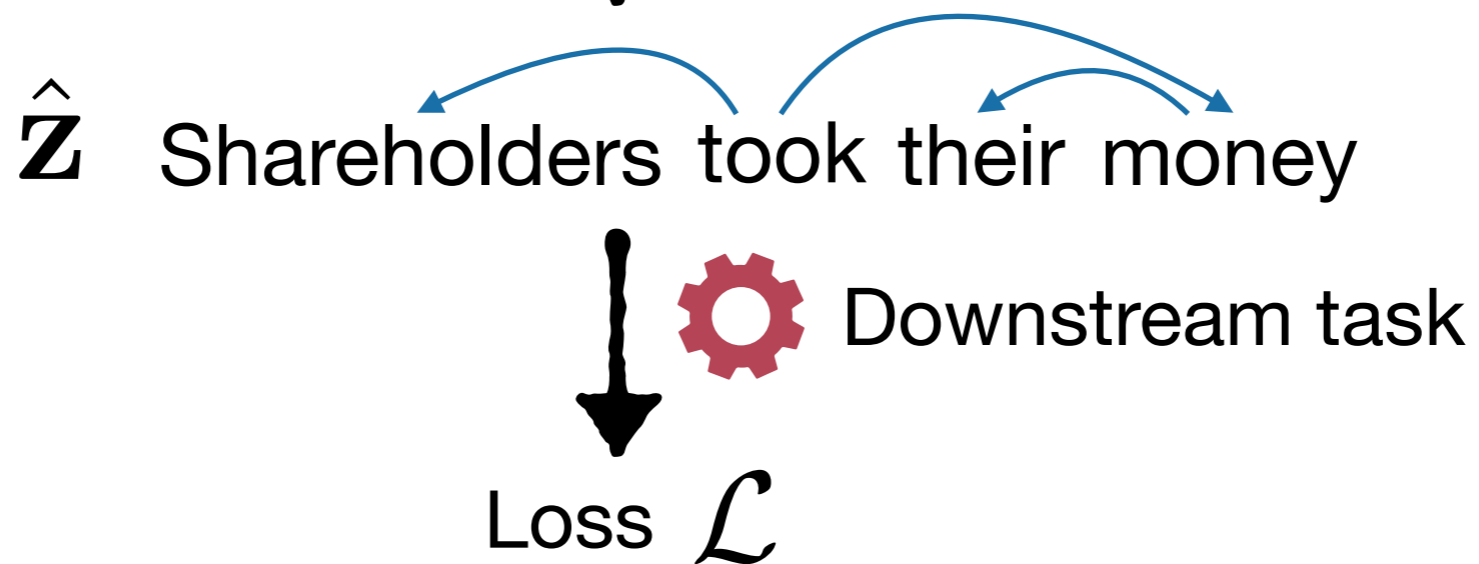


Outline

- ❖ Background: structured prediction as linear programs
- ❖ **Method: SPIGOT algorithm**
- ❖ Experiments

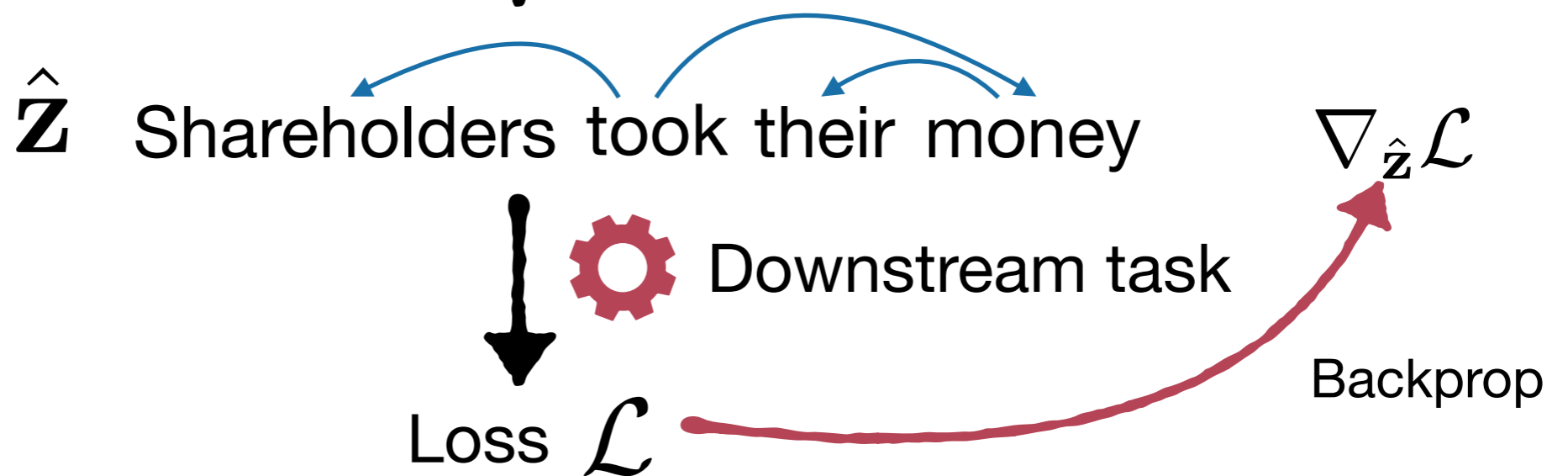
Backprop

$$\begin{array}{l} \arg \max \mathbf{z}^\top \\ \text{s.t. } \mathbf{Z} \text{ forms a tree} \end{array} \quad \begin{bmatrix} s_\theta \text{ (their money)} \\ s_\theta \text{ (took their)} \\ s_\theta \text{ (took money)} \\ \vdots \\ s_\theta \text{ (their took)} \end{bmatrix} \quad \textcircled{\nabla_\theta \mathcal{L}}$$

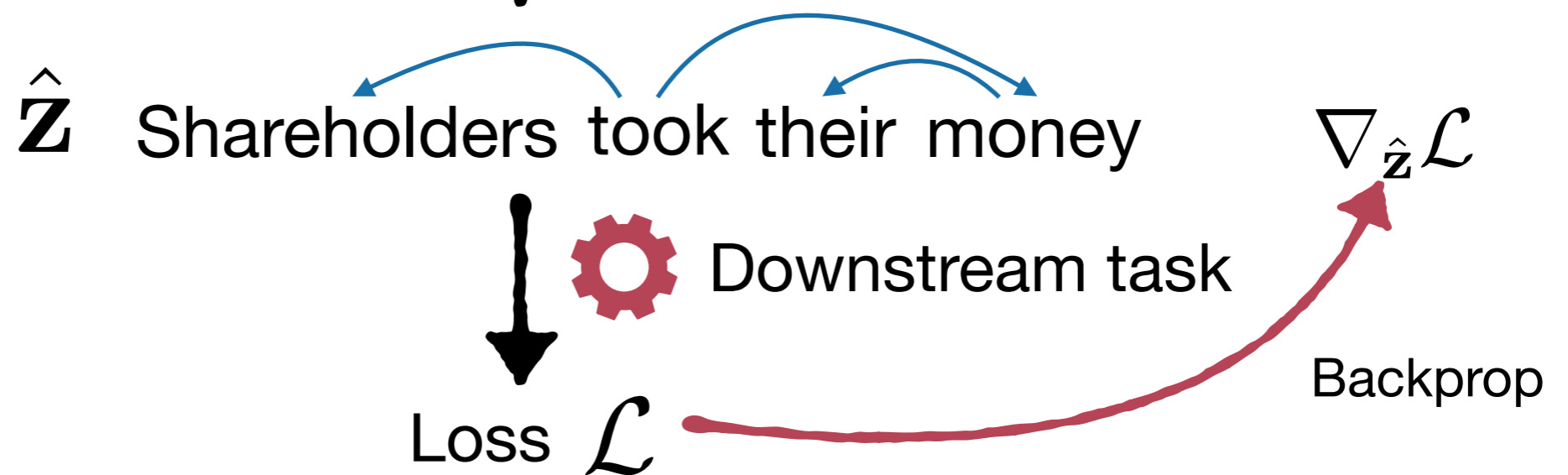
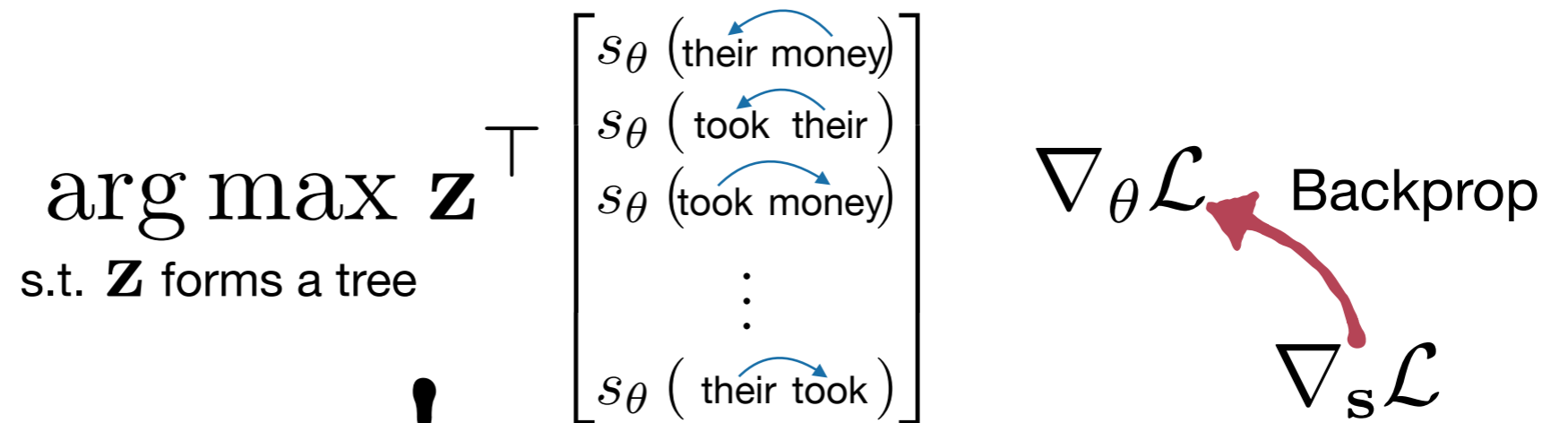


Backprop

$$\begin{array}{l} \arg \max \mathbf{z}^\top \\ \text{s.t. } \mathbf{Z} \text{ forms a tree} \end{array} \quad \begin{bmatrix} s_\theta \text{ (their money)} \\ s_\theta \text{ (took their)} \\ s_\theta \text{ (took money)} \\ \vdots \\ s_\theta \text{ (their took)} \end{bmatrix} \quad \nabla_\theta \mathcal{L}$$



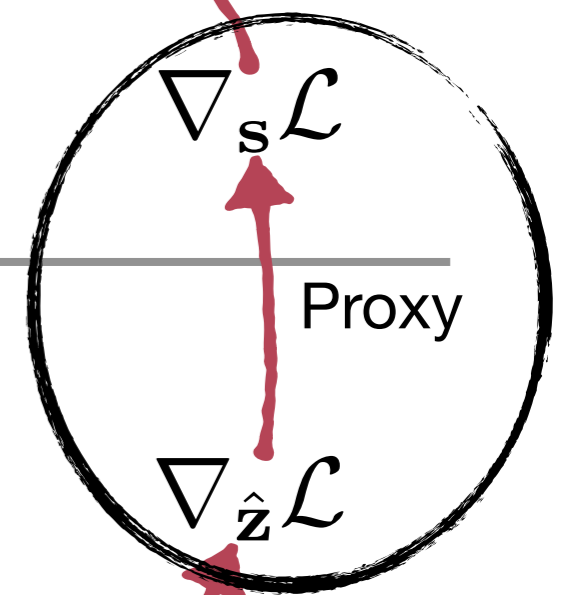
Backprop



Backprop

$$\begin{aligned} & \arg \max \mathbf{z}^\top \\ & \text{s.t. } \mathbf{Z} \text{ forms a tree} \end{aligned} \quad \begin{bmatrix} s_\theta \text{ (their money)} \\ s_\theta \text{ (took their)} \\ s_\theta \text{ (took money)} \\ \vdots \\ s_\theta \text{ (their took)} \end{bmatrix}$$

$\nabla_\theta \mathcal{L}$ Backprop



$\hat{\mathbf{z}}$ Shareholders took their money



Downstream task

Loss \mathcal{L}

Backprop

Backprop

We have: $\nabla_{\hat{z}} \mathcal{L}$

We need: $\nabla_{\mathbf{s}} \mathcal{L}$

Backprop

We have: $\nabla_{\hat{\mathbf{z}}}\mathcal{L}$

We need: $\nabla_{\mathbf{s}}\mathcal{L}$

Leibniz, 1676

$$\nabla_{\mathbf{s}}\mathcal{L} = \text{“J”} \nabla_{\hat{\mathbf{z}}}\mathcal{L} \quad \times$$

Backprop

We have: $\nabla_{\hat{\mathbf{z}}} \mathcal{L}$

We need: $\nabla_{\mathbf{s}} \mathcal{L}$

Leibniz, 1676

$$\nabla_{\mathbf{s}} \mathcal{L} = \text{“J”} \nabla_{\hat{\mathbf{z}}} \mathcal{L} \quad \times$$

$$\hat{\mathbf{z}} = \underset{\text{s.t. } \mathbf{Z} \text{ forms a tree}}{\text{arg max}} \mathbf{Z}^T \mathbf{s}_\theta$$

Jacobian not defined

Backprop

We have: $\nabla_{\hat{\mathbf{z}}}\mathcal{L}$

We need: $\nabla_{\mathbf{s}}\mathcal{L}$

Leibniz, 1676

$$\nabla_{\mathbf{s}}\mathcal{L} = \text{“J”} \nabla_{\hat{\mathbf{z}}}\mathcal{L} \quad \times$$

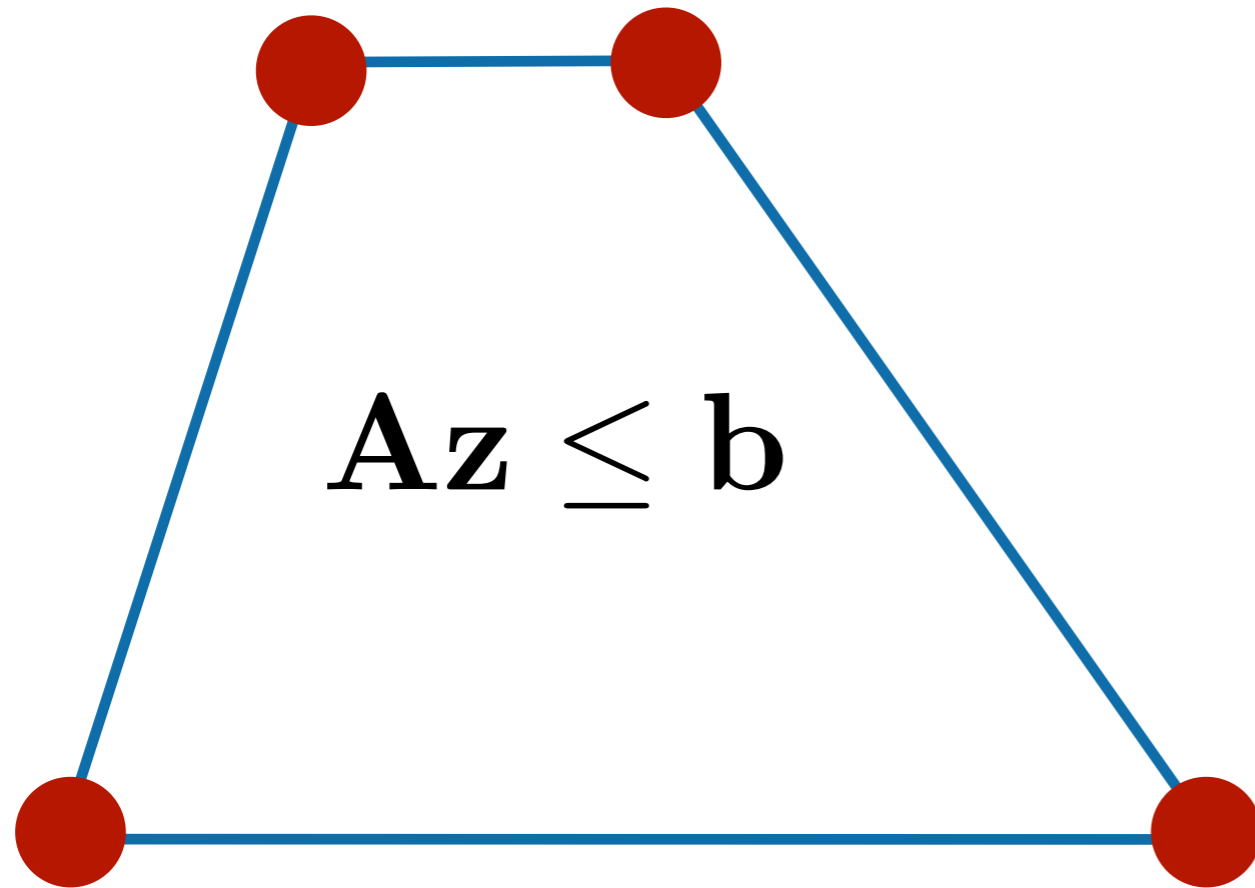
Straight-through Estimator (STE)

Hinton, 2012; Bengio et al., 2013

$$\nabla_{\mathbf{s}}\mathcal{L} \triangleq \nabla_{\hat{\mathbf{z}}}\mathcal{L}$$

Some Geometry...

Straight-through Estimator (STE): $\nabla_{\mathbf{s}} \mathcal{L} \triangleq \nabla_{\hat{\mathbf{z}}} \mathcal{L}$



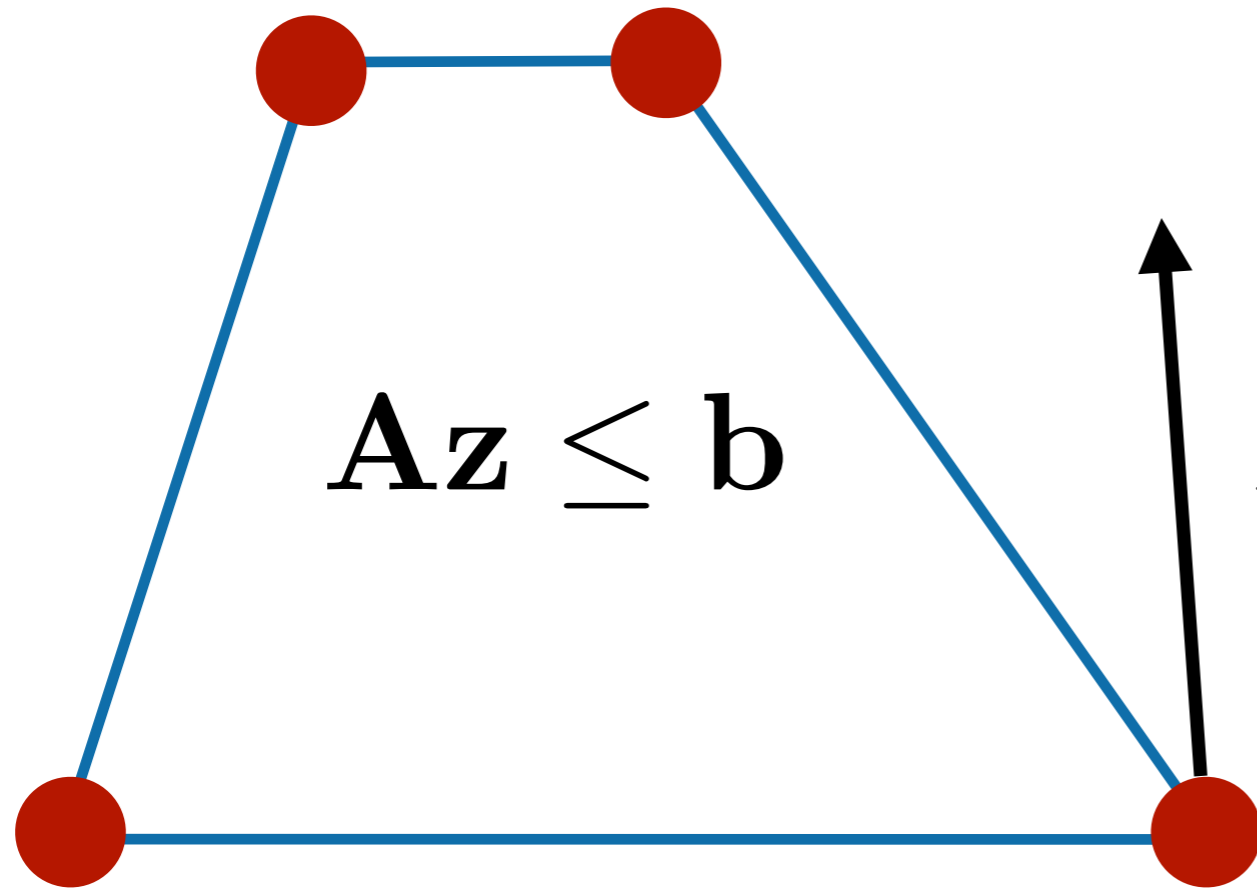
$$\hat{\mathbf{z}} = [1, 0, 1, \dots, 0]^\top$$

Shareholders took their money

Three blue curved arrows point from the text 'Shareholders took their money' to the 1st, 3rd, and 5th elements of the vector $\hat{\mathbf{z}}$.

Some Geometry...

Straight-through Estimator (STE): $\nabla_{\mathbf{s}} \mathcal{L} \triangleq \nabla_{\hat{\mathbf{z}}} \mathcal{L}$



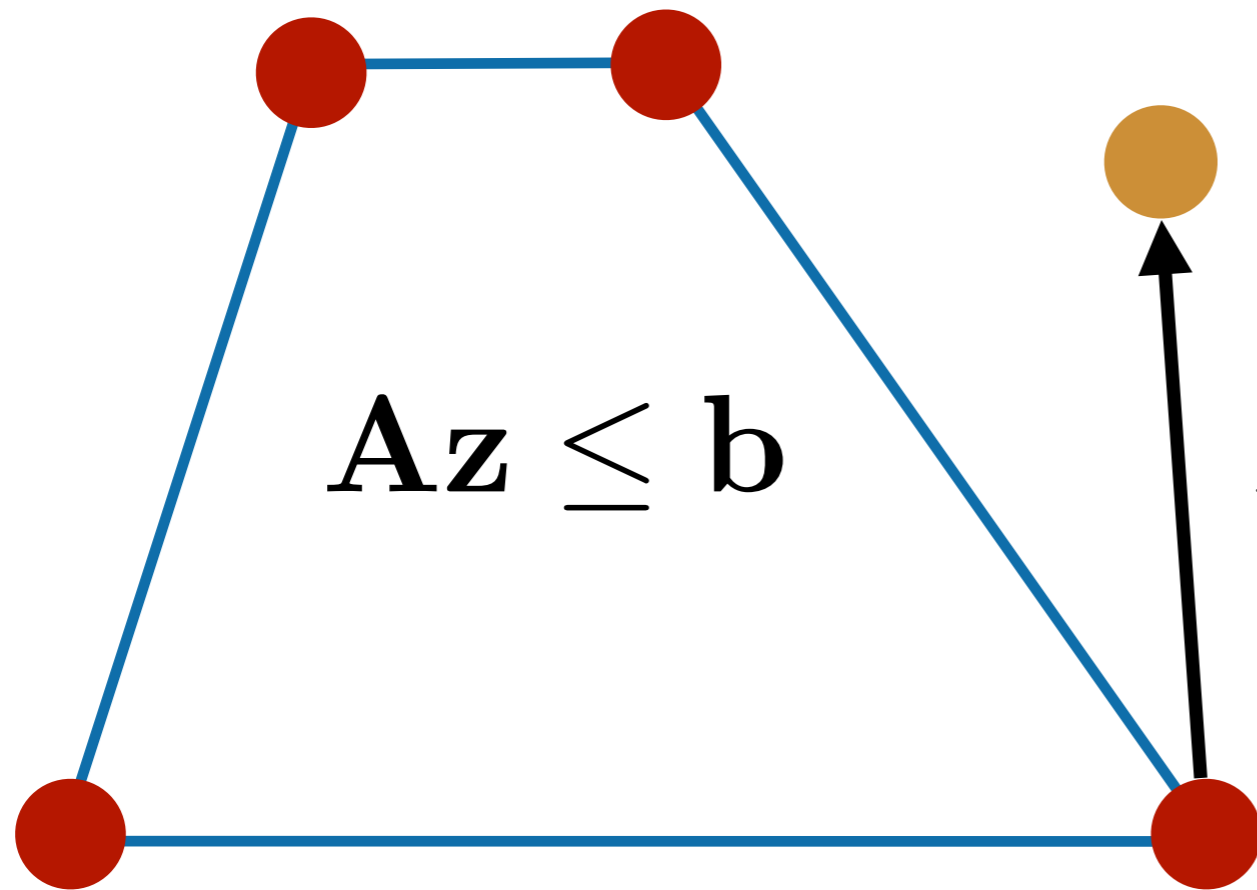
$$-\nabla_{\hat{\mathbf{z}}} \mathcal{L} = [-0.3, 0.5, 0.4, \dots, 0.2]$$

$$\hat{\mathbf{z}} = [1, 0, 1, \dots, 0]^{\top}$$

Shareholders took their money

Some Geometry...

Straight-through Estimator (STE): $\nabla_{\mathbf{s}} \mathcal{L} \triangleq \nabla_{\hat{\mathbf{z}}} \mathcal{L}$



$$\mathbf{p} = \hat{\mathbf{z}} - \nabla_{\hat{\mathbf{z}}} \mathcal{L}$$

Shareholders took their money

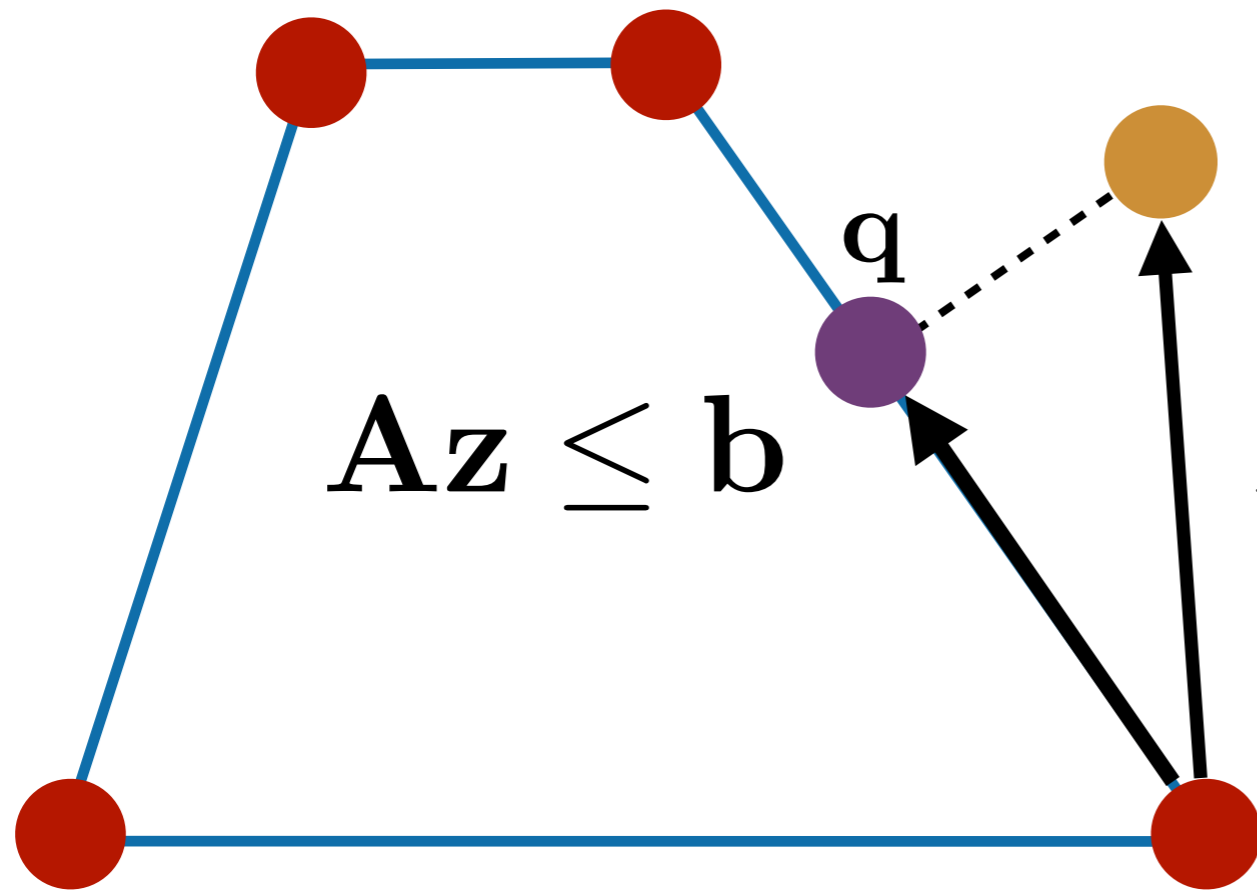
$$-\nabla_{\hat{\mathbf{z}}} \mathcal{L} = [-0.3, 0.5, 0.4, \dots, 0.2]$$

$$\hat{\mathbf{z}} = [1, 0, 1, \dots, 0]^{\top}$$

Shareholders took their money

Some Geometry...

SPIGOT



$$p = \hat{z} - \nabla_{\hat{z}} \mathcal{L}$$

Shareholders took their money

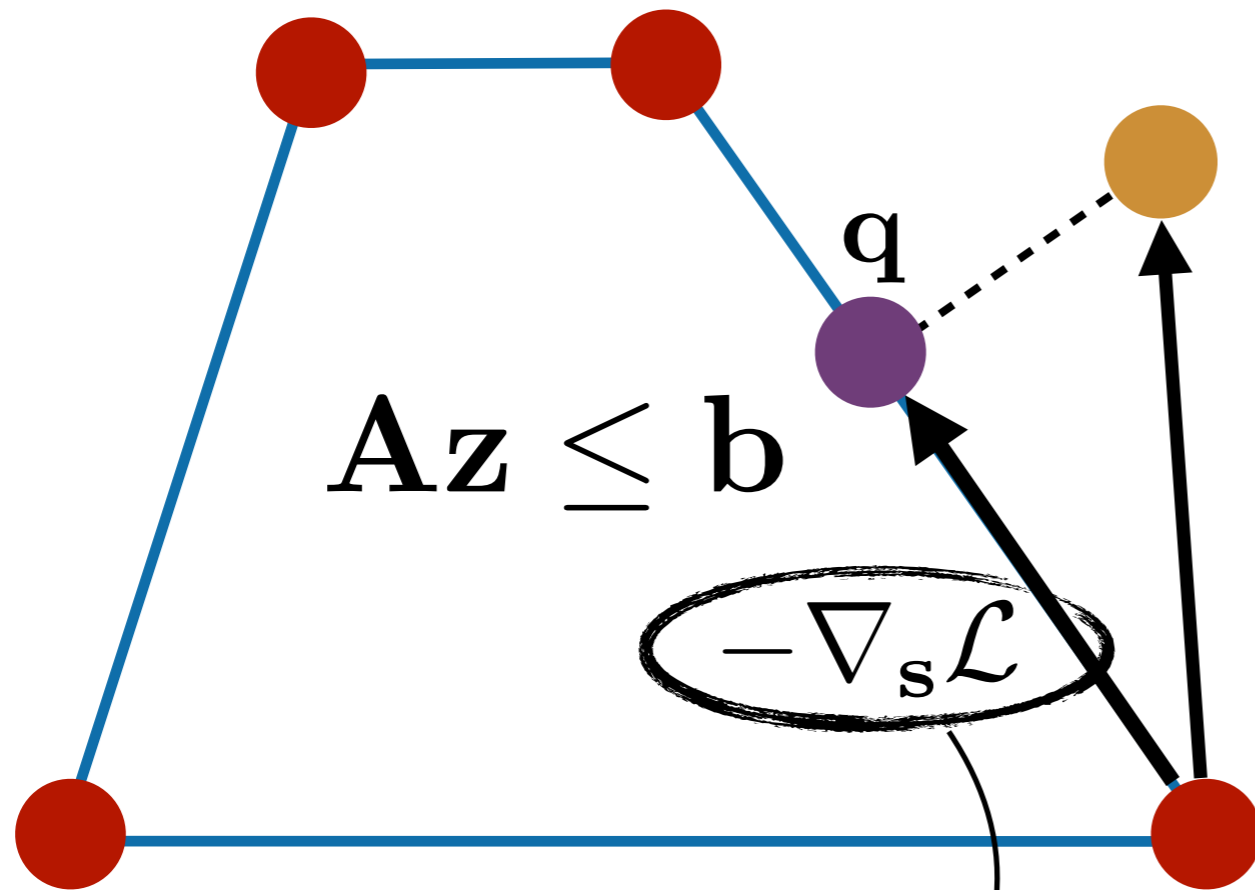
$$-\nabla_{\hat{z}} \mathcal{L} = [-0.3, 0.5, 0.4, \dots, 0.2]$$

$$\hat{z} = [1, 0, 1, \dots, 0]^T$$

Shareholders took their money

Some Geometry...

SPIGOT



$$\mathbf{A}\mathbf{z} \leq \mathbf{b}$$

$$-\nabla_s \mathcal{L}$$

$$\begin{aligned} \mathbf{p} &= \hat{\mathbf{z}} - \nabla_{\hat{\mathbf{z}}} \mathcal{L} \\ \mathbf{q} &= \text{proj}(\mathbf{p}) \\ \nabla_s \mathcal{L} &\triangleq \hat{\mathbf{z}} - \mathbf{q} \end{aligned}$$

$$\mathbf{p} = \hat{\mathbf{z}} - \nabla_{\hat{\mathbf{z}}} \mathcal{L}$$

Shareholders took their money

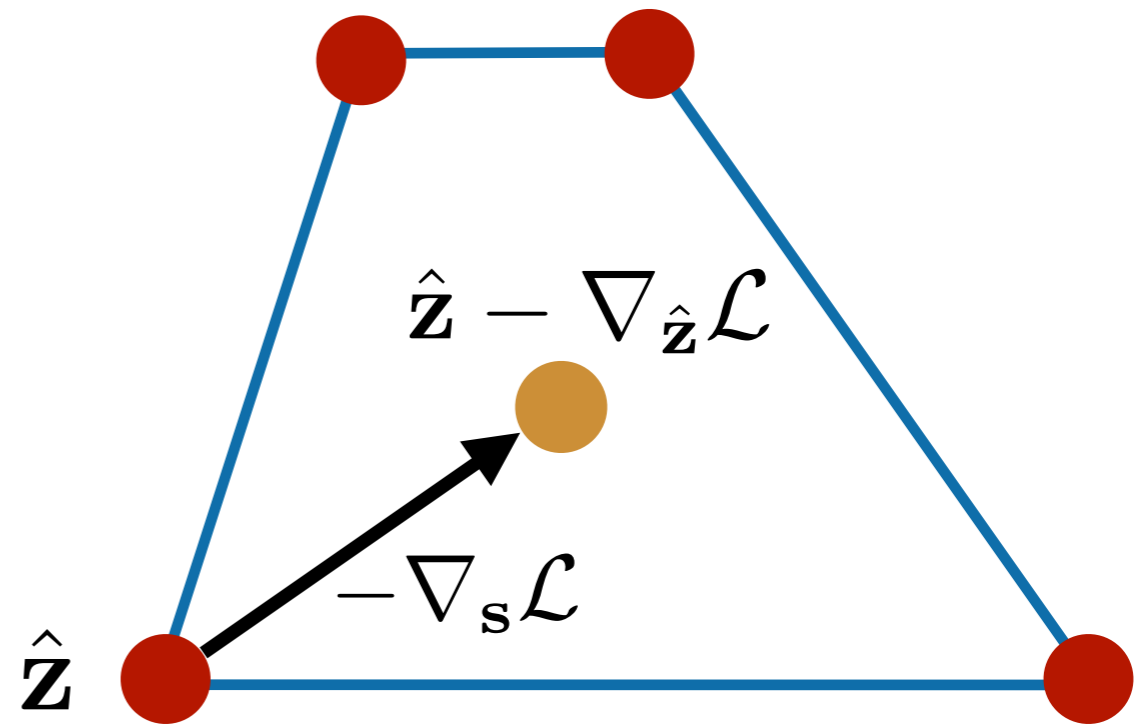
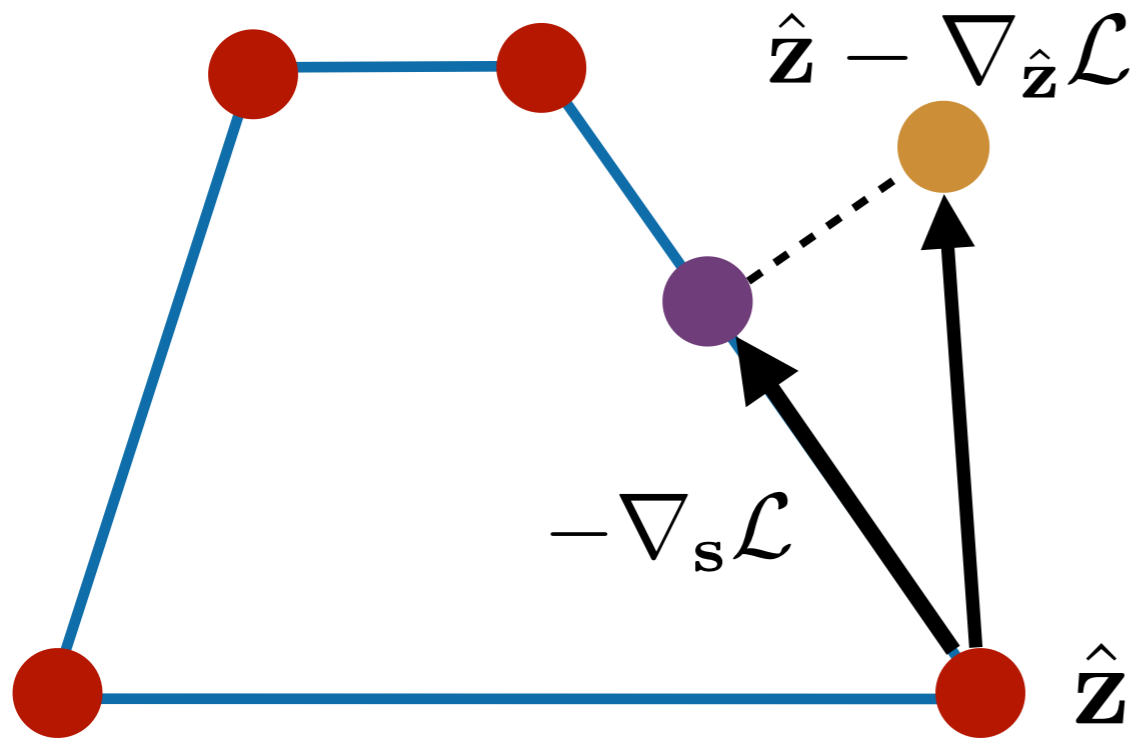
$$-\nabla_{\hat{\mathbf{z}}} \mathcal{L} = [-0.3, 0.5, 0.4, \dots, 0.2]$$

$$\hat{\mathbf{z}} = [1, 0, 1, \dots, 0]^\top$$

Shareholders took their money

Some Geometry...

SPIGOT



Algorithm

Input

Shareholders took their money

 Parser θ

$$\arg \max_{\mathbf{z}} \mathbf{z}^T \begin{bmatrix} s_{\theta}(\text{their money}) \\ s_{\theta}(\text{took their}) \\ s_{\theta}(\text{took money}) \\ \vdots \\ s_{\theta}(\text{their took}) \end{bmatrix}$$

s.t. \mathbf{z} forms a tree

$\hat{\mathbf{z}}$ Shareholders took their money

Algorithm

Input

Shareholders took their money

 Parser θ

$$\arg \max_{\mathbf{z}} \mathbf{z}^T \begin{bmatrix} s_{\theta}(\text{their money}) \\ s_{\theta}(\text{took their}) \\ s_{\theta}(\text{took money}) \\ \vdots \\ s_{\theta}(\text{their took}) \end{bmatrix}$$

s.t. \mathbf{z} forms a tree

$\hat{\mathbf{z}}$ Shareholders took their money

 Downstream task ϕ

Loss \mathcal{L}

Algorithm

Input

Shareholders took their money

 Parser θ

$$\arg \max_{\mathbf{z}} \mathbf{z}^{\top} \begin{bmatrix} s_{\theta}(\text{their money}) \\ s_{\theta}(\text{took their}) \\ s_{\theta}(\text{took money}) \\ \vdots \\ s_{\theta}(\text{their took}) \end{bmatrix}$$

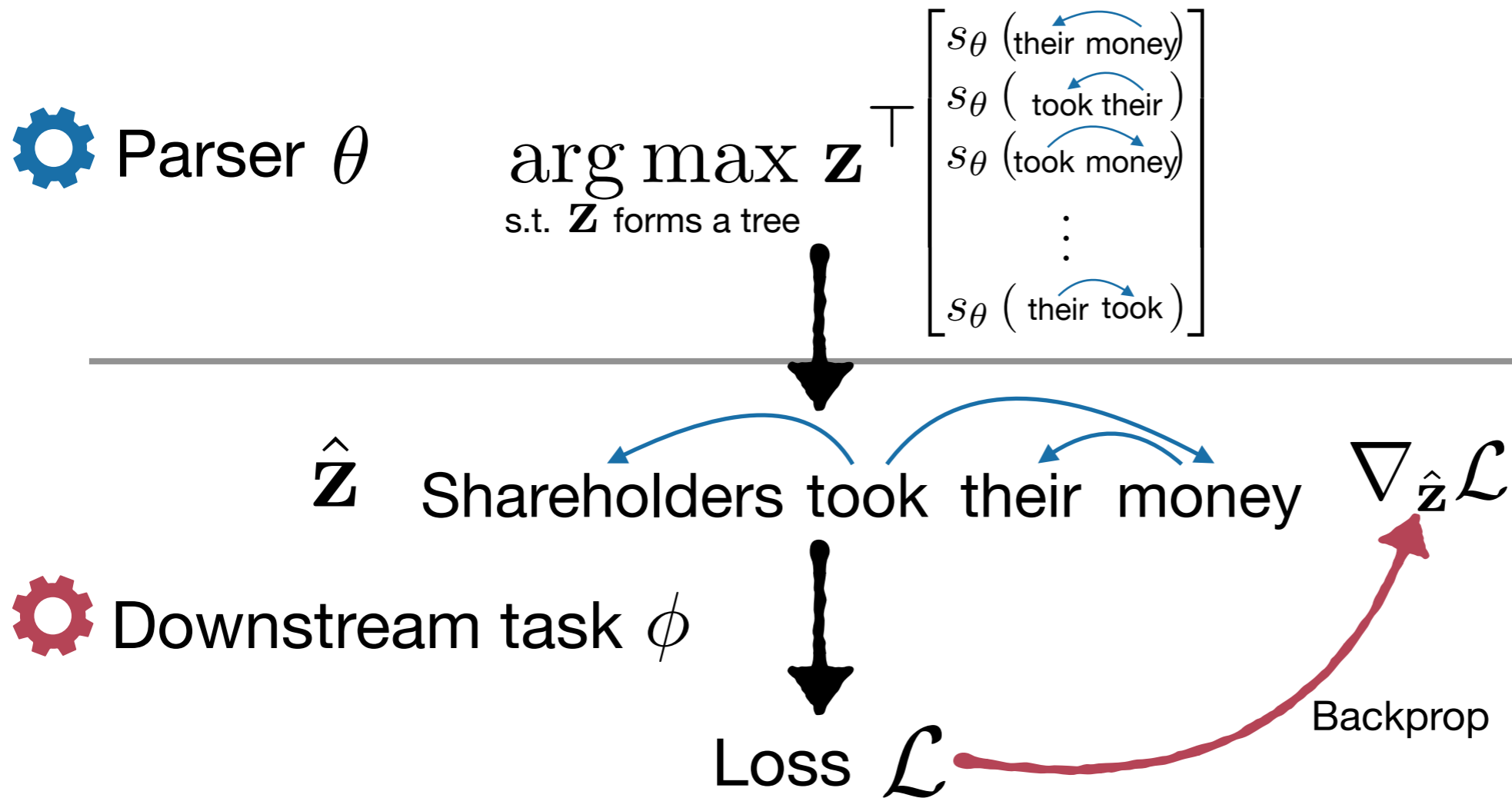
s.t. \mathbf{z} forms a tree

$\hat{\mathbf{z}}$ Shareholders took their money $\nabla_{\hat{\mathbf{z}}} \mathcal{L}$

 Downstream task ϕ

Loss \mathcal{L}

Backprop



Algorithm

Input

Shareholders took their money

 Parser θ

$$\arg \max_{\mathbf{Z}} \mathbf{z}^T \begin{bmatrix} s_{\theta}(\text{their money}) \\ s_{\theta}(\text{took their}) \\ s_{\theta}(\text{took money}) \\ \vdots \\ s_{\theta}(\text{their took}) \end{bmatrix}$$

s.t. \mathbf{Z} forms a tree

$\hat{\mathbf{Z}}$ Shareholders took their money

 Downstream task ϕ

Loss \mathcal{L}

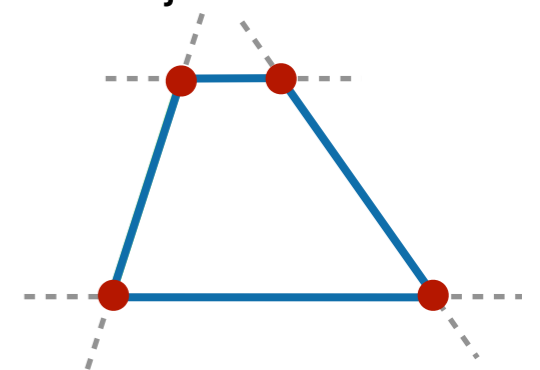
$\nabla_{\mathbf{s}} \mathcal{L}$

$\nabla_{\hat{\mathbf{z}}} \mathcal{L}$

Backprop

$\mathbf{p} = \hat{\mathbf{z}} - \nabla_{\hat{\mathbf{z}}} \mathcal{L}$
 $\mathbf{q} = \text{proj}(\mathbf{p})$
 $\nabla_{\mathbf{s}} \mathcal{L} \triangleq \hat{\mathbf{z}} - \mathbf{q}$

Project onto



Algorithm

Input

Shareholders took their money

 Parser θ

$$\arg \max_{\mathbf{Z}} \mathbf{z}^\top \begin{bmatrix} s_\theta(\text{their money}) \\ s_\theta(\text{took their}) \\ s_\theta(\text{took money}) \\ \vdots \\ s_\theta(\text{their took}) \end{bmatrix}$$

s.t. \mathbf{Z} forms a tree

$$\nabla_{\theta} \mathcal{L}$$

Backprop

$$\nabla_{\mathbf{s}} \mathcal{L}$$

$$\mathbf{p} = \hat{\mathbf{z}} - \nabla_{\hat{\mathbf{z}}} \mathcal{L}$$
$$\mathbf{q} = \text{proj}(\mathbf{p})$$
$$\nabla_{\mathbf{s}} \mathcal{L} \triangleq \hat{\mathbf{z}} - \mathbf{q}$$

$\hat{\mathbf{Z}}$ Shareholders took their money

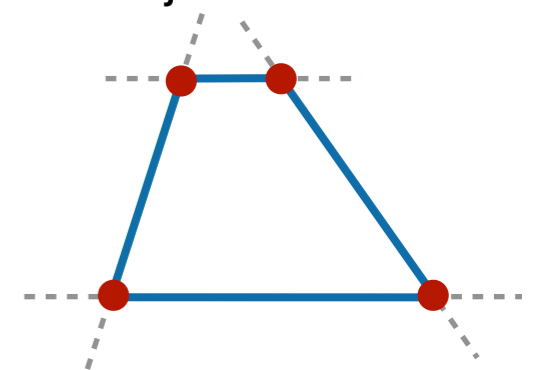
$$\nabla_{\hat{\mathbf{z}}} \mathcal{L}$$

 Downstream task ϕ

Loss \mathcal{L}

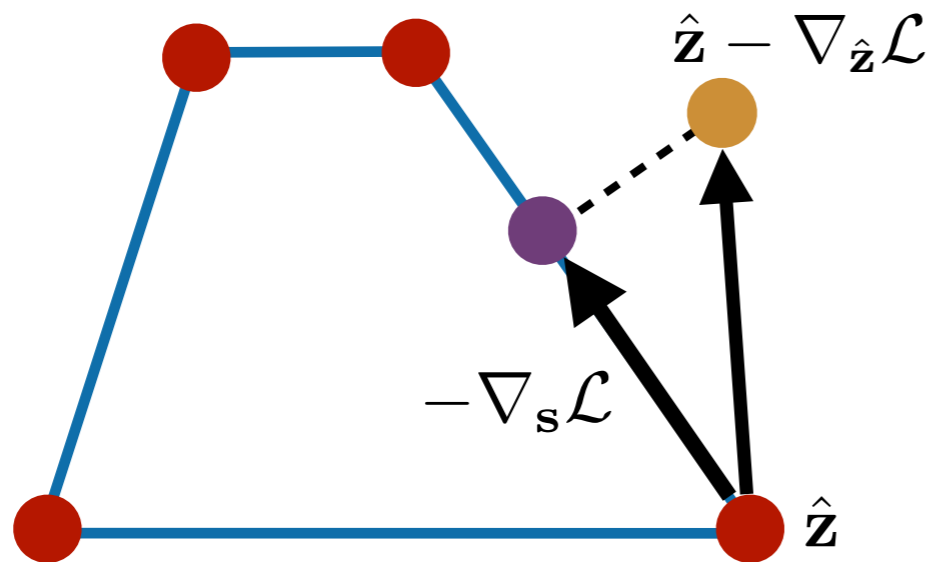
Backprop

Project onto

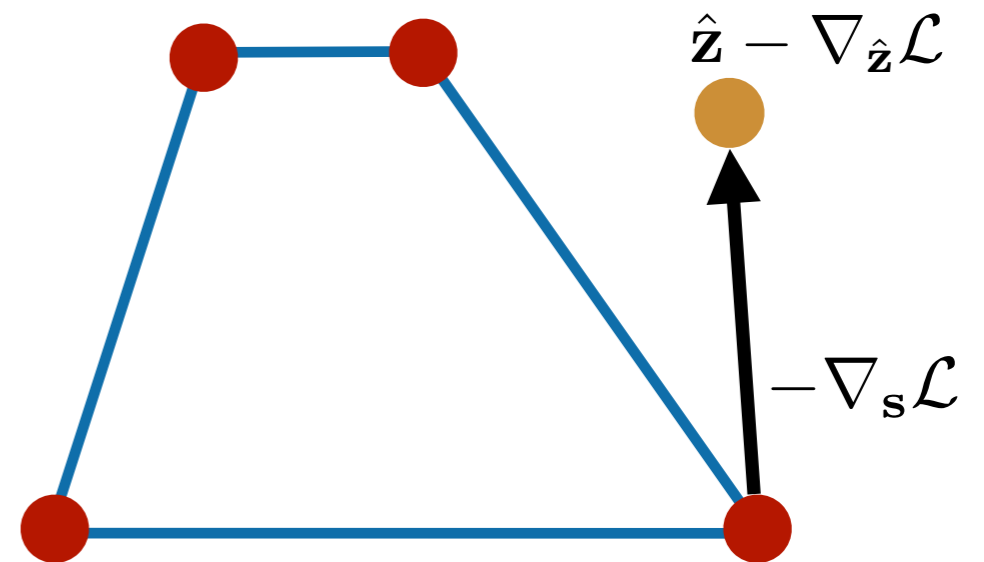


Connections to Related Work

SPIGOT



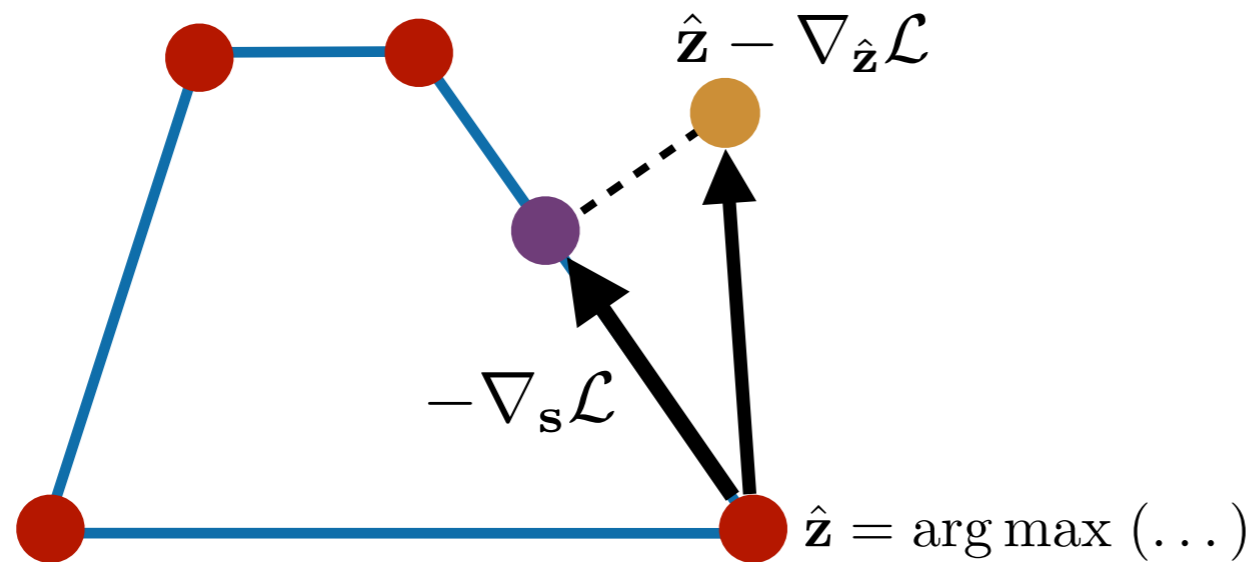
STE



	Pipeline	STE	Structured Att.	SPIGOT
Hard decision on $\hat{\mathbf{z}}$	✓	✓		✓
Backprop		✓	✓	✓
Marginal			✓	
Projection				✓

Connections to Related Work

SPIGOT



Structured Attention

$$\hat{\mathbf{z}} = \text{softmax}(\dots)$$

	Pipeline	STE	Structured Att.	SPIGOT
Hard decision on $\hat{\mathbf{z}}$	✓	✓		✓
Backprop		✓	✓	✓
Marginal			✓	
Projection				✓

Applications

Joint learning

Swayamdipta et al., 2016

Training data



Shareholders took their money



Shareholders took their money

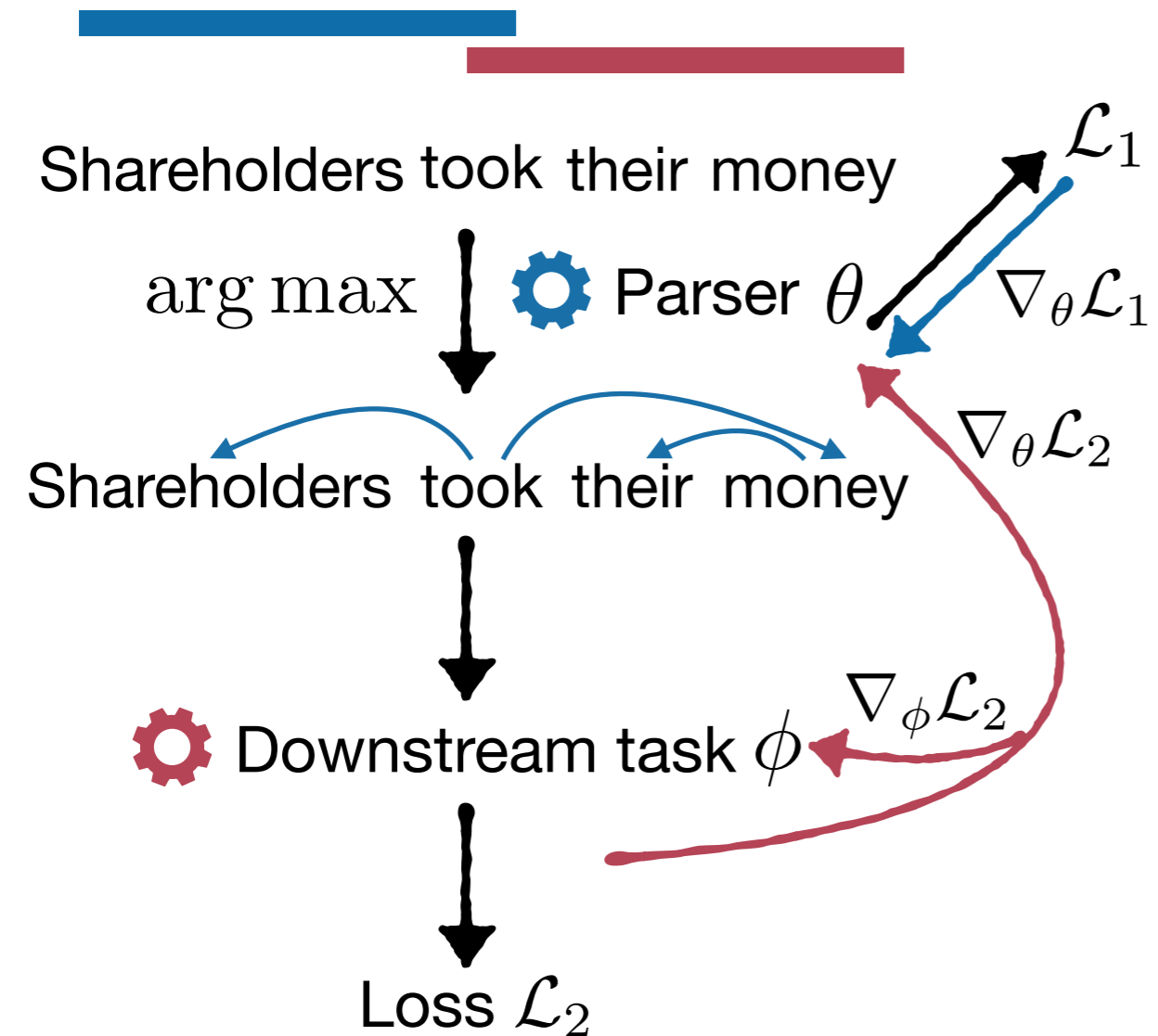


Applications

Joint learning

Swayamdipta et al., 2016

Training data

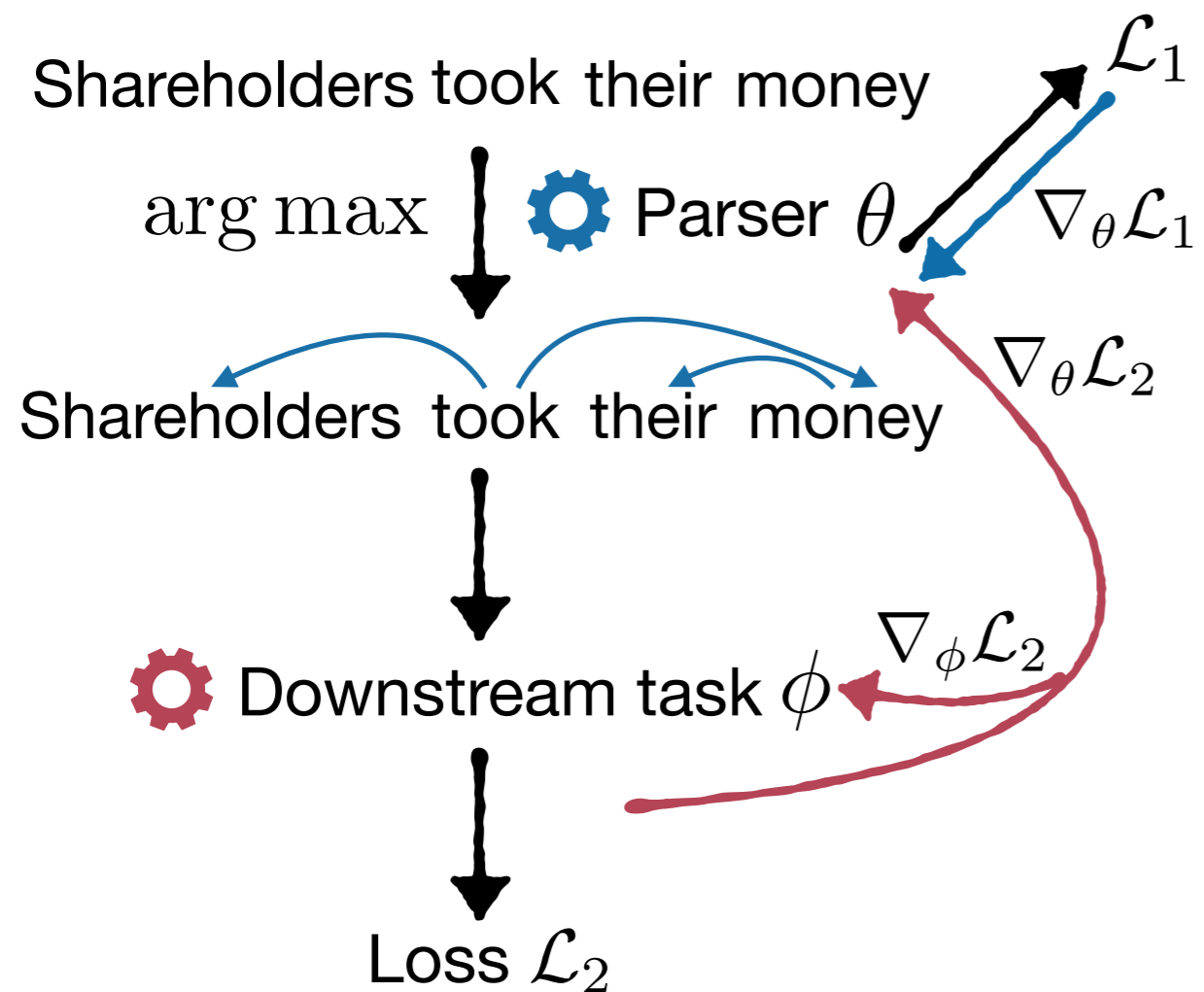


Applications

Joint learning

Swayamdipta et al., 2016

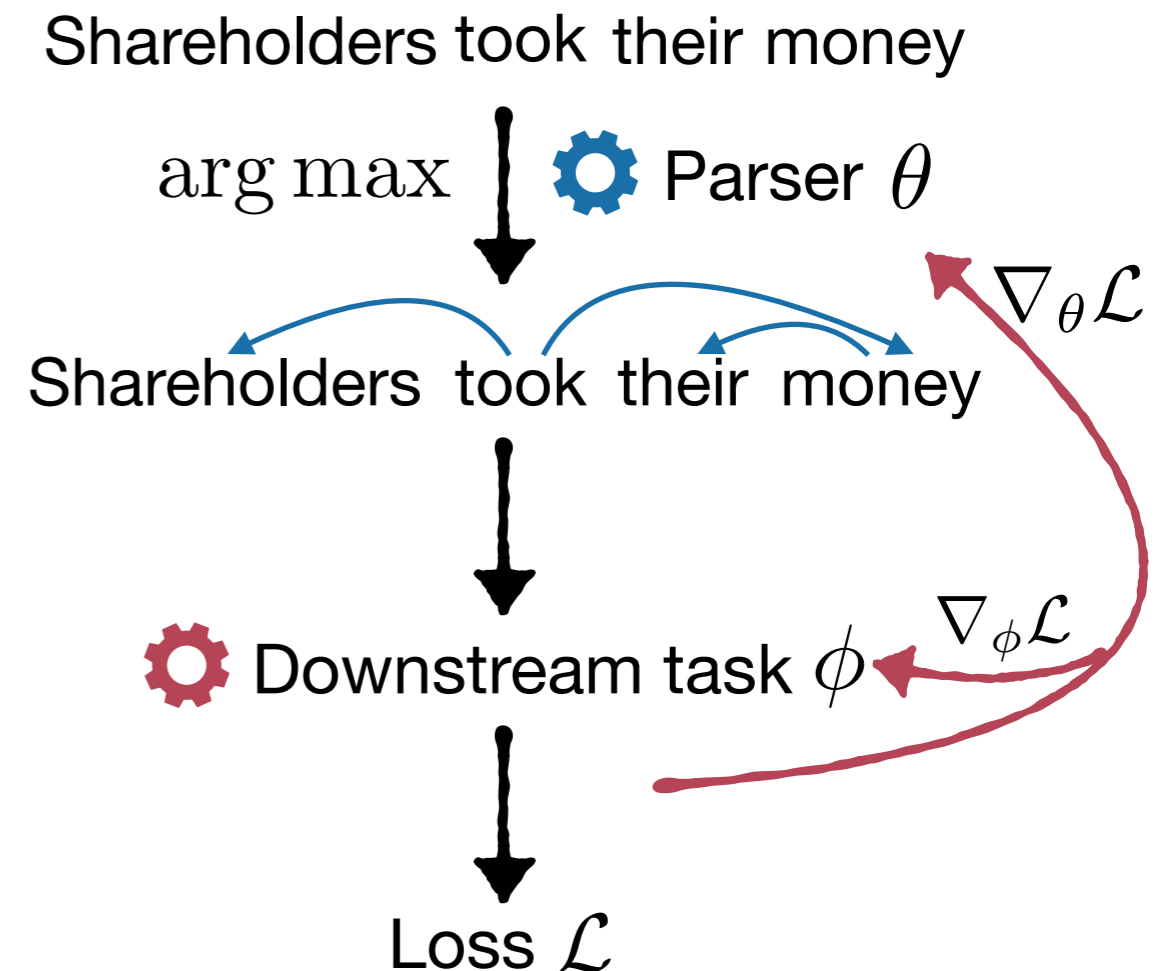
Training data



Induce latent structures

Yogatama et al., 2017; Williams et al., 2017

Training data



Outline

- ❖ Background: structured prediction as linear programs
- ❖ Method: SPIGOT algorithm
- ❖ **Experiments**

Experiments: Syntactic-then-semantic Parsing

Input


Shareholders took their money

arg max ↓  Syntactic Parser θ

Syntactic tree

Shareholders took their money

Semantic graph

↓  Semantic Parser ϕ

Shareholders took their money

Experiments: Syntactic-then-semantic Parsing

Input

Shareholders took their money

BiLSTM + MLP
Kiperwasser and Goldberg, 2016

Eisner Algorithm
Eisner, 1996



Syntactic tree

Shareholders took their money

Semantic graph

Shareholders took their money

Experiments: Syntactic-then-semantic Parsing

Input

Shareholders took their money

BiLSTM + MLP

Kiperwasser and Goldberg, 2016

Eisner Algorithm

Eisner, 1996

arg max



Syntactic Parser



Syntactic tree



NeurboParser

Peng et al., 2017

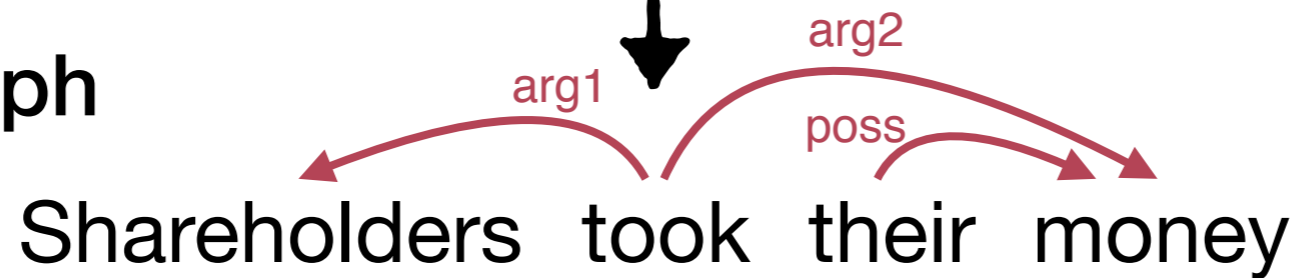
Concat head token embedding



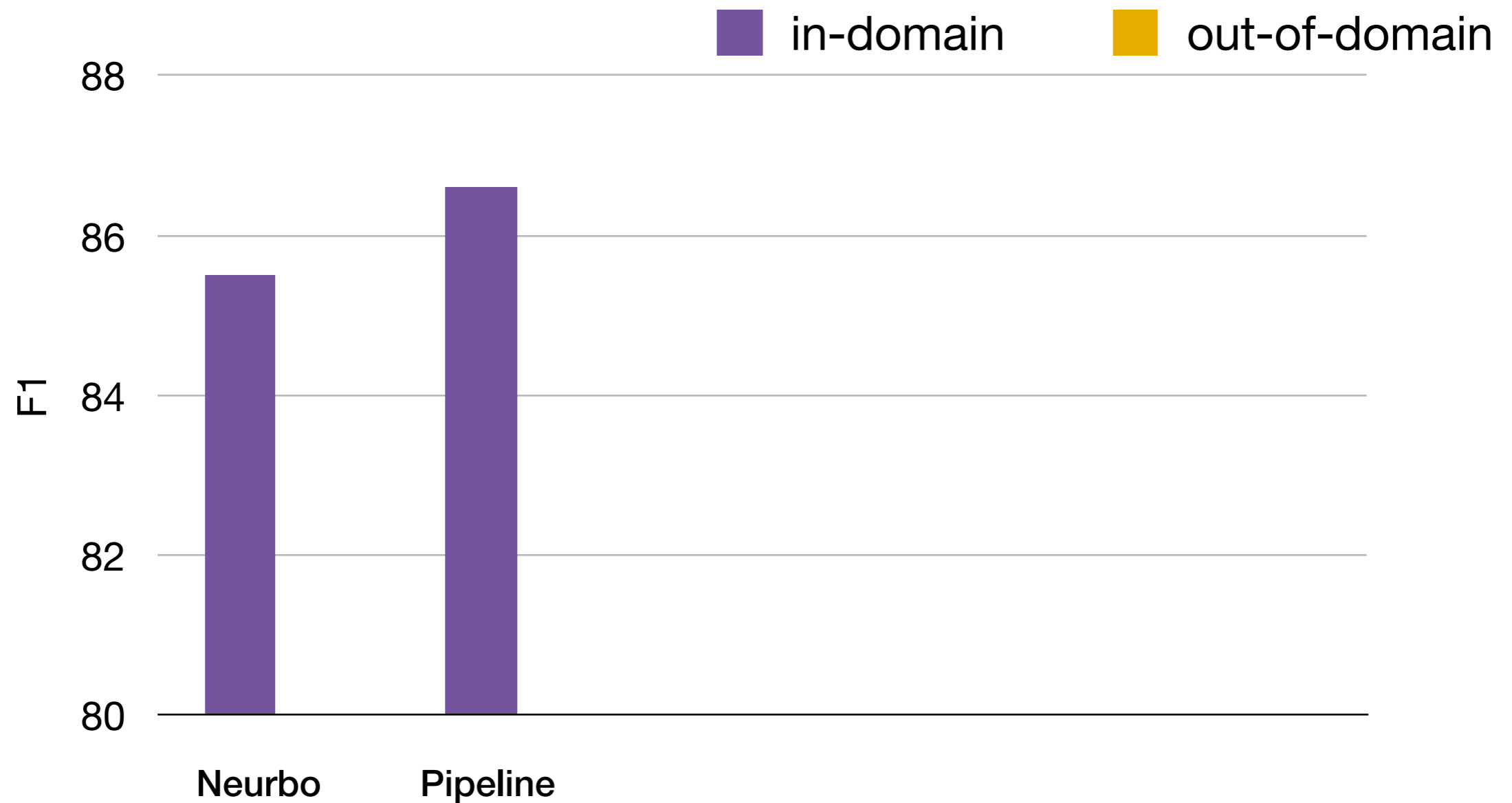
Semantic Parser



Semantic graph

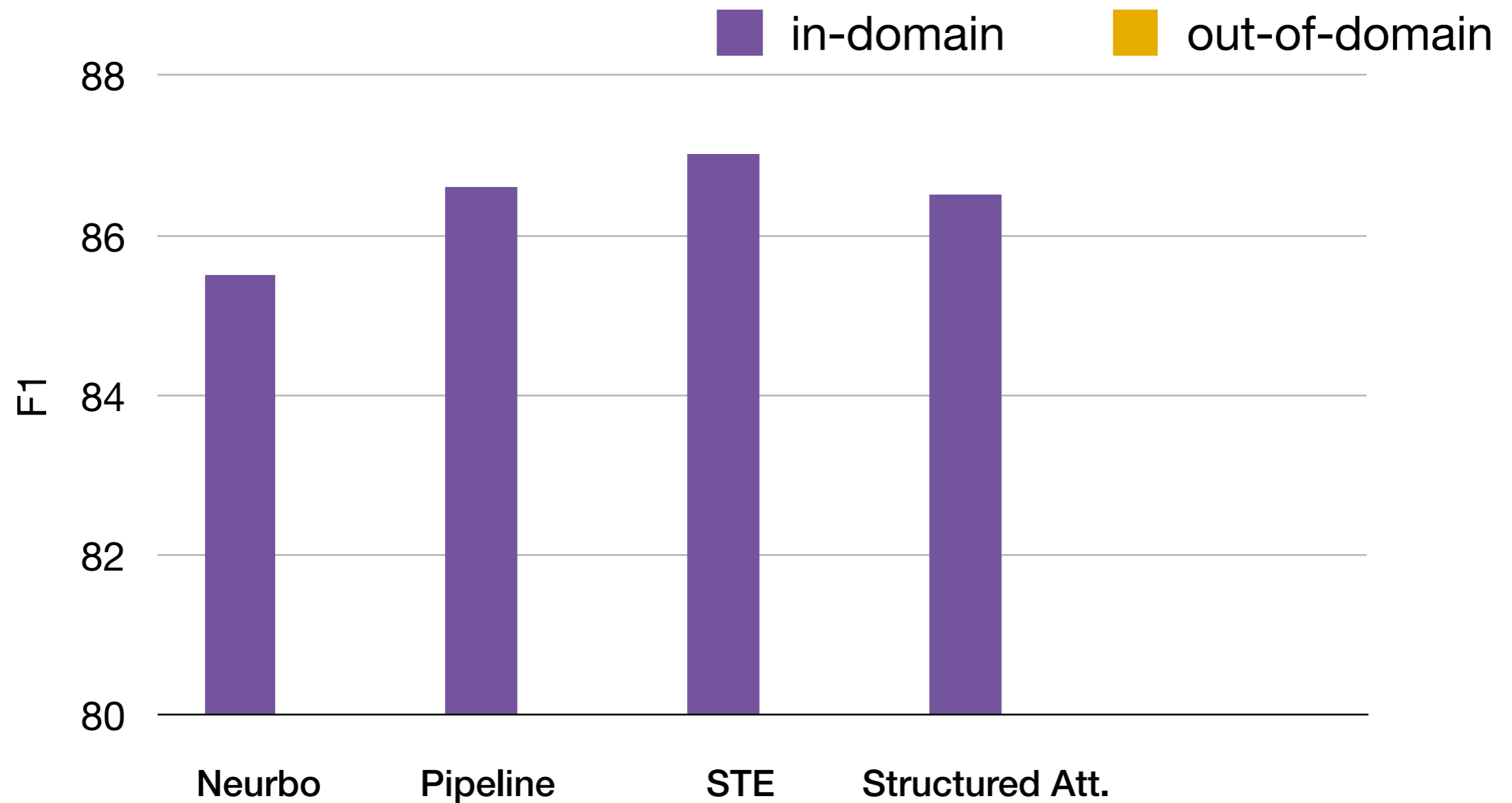


SemEval '15. Micro-averaged labeled F_1



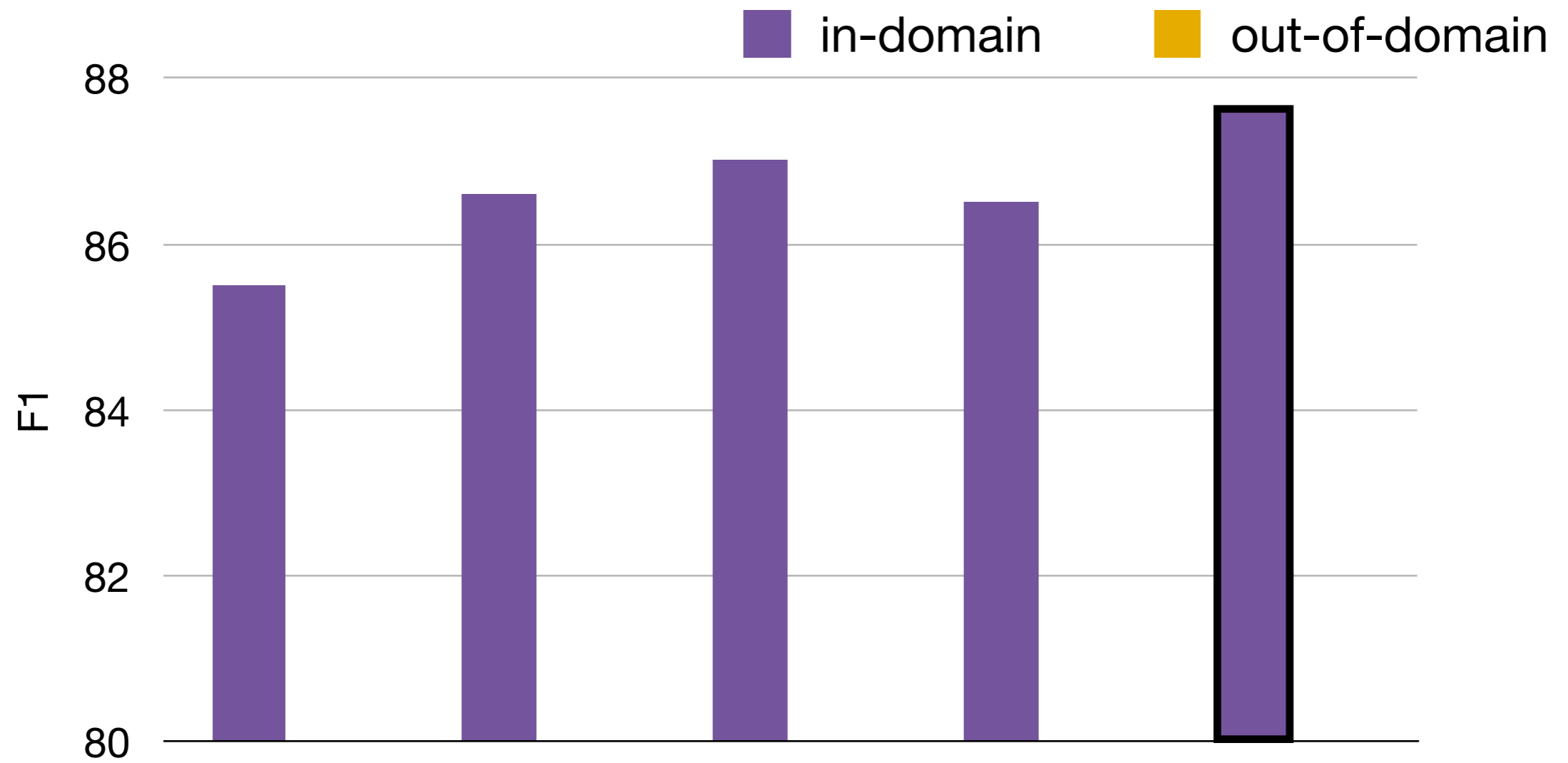
	Neurbo	Pipeline
Syntax		✓
Backprop	N/A	
Hard decision \hat{Z}	N/A	✓
Projection	N/A	

SemEval '15. Micro-averaged labeled F_1



	Neurbo	Pipeline	STE	Structured Att.
Syntax		✓	✓	✓
Backprop	N/A		✓	✓
Hard decision \hat{Z}	N/A	✓	✓	
Projection	N/A			

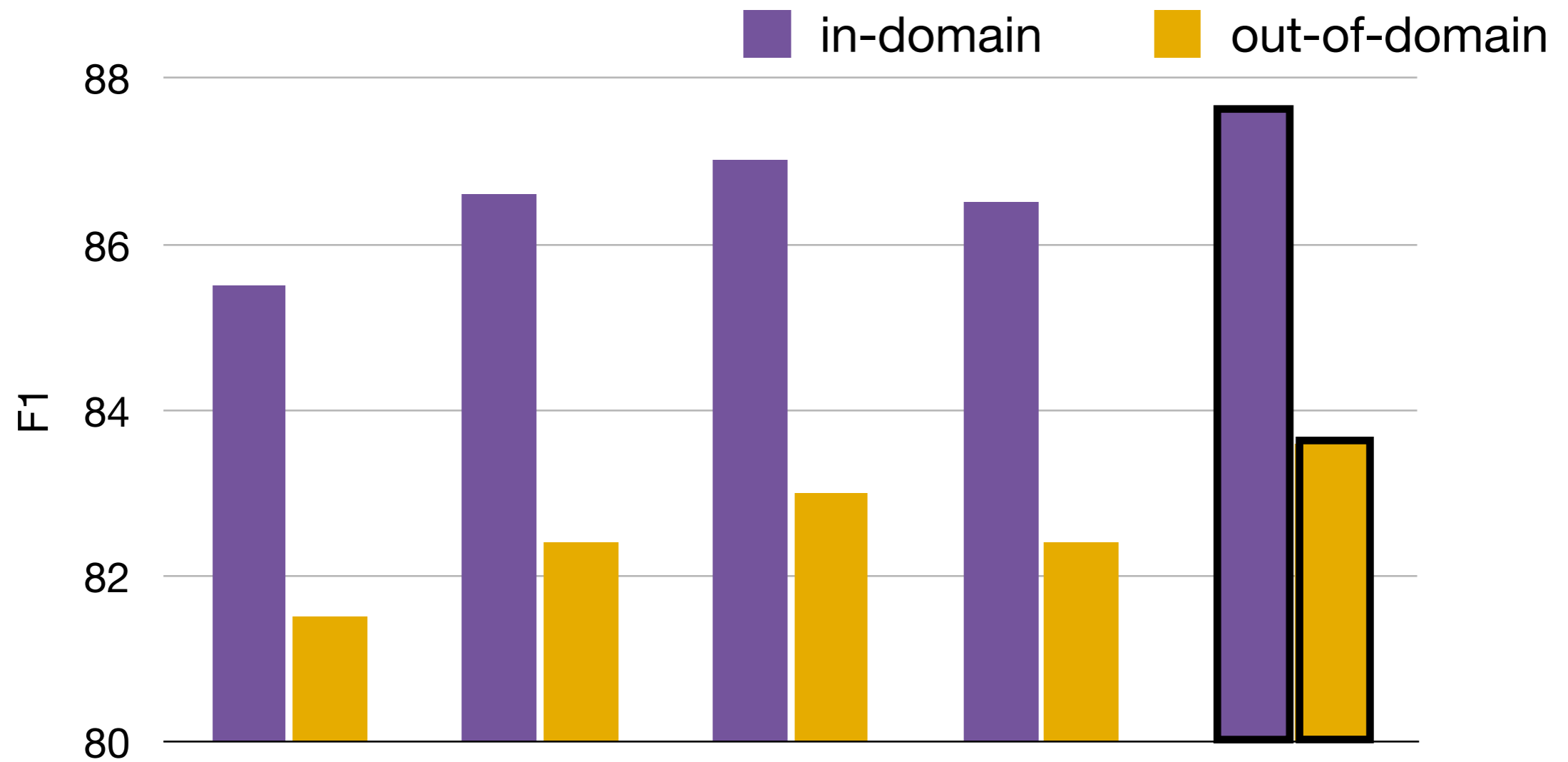
SemEval '15. Micro-averaged labeled F_1



	Neurbo	Pipeline	STE	Structured Att.	SPIGOT
Syntax		✓	✓	✓	✓
Backprop	N/A		✓	✓	✓
Hard decision \hat{Z}	N/A	✓	✓		✓
Projection	N/A				✓

Neurbo: Peng et al., 2017

SemEval '15. Micro-averaged labeled F_1



	Neurbo	Pipeline	STE	Structured Att.	SPIGOT
Syntax		✓	✓	✓	✓
Backprop	N/A		✓	✓	✓
Hard decision \hat{Z}	N/A	✓	✓		✓
Projection	N/A				✓

Neurbo: Peng et al., 2017

Semantic Parsing for Sentiment Classification


Input

Shareholders took their money

arg max  Semantic Parser θ

Semantic graph

Shareholders took their money



 Classifier ϕ

Positive? Negative?

Semantic Parsing for Sentiment Classification

Input

Shareholders took their money

NeurboParser
Peng et al., 2017

AD³
Martins et al., 2011

arg max



Semantic Parser



Semantic graph

Shareholders
took: arg1

arg1

took

...

arg2

poss

their money

...

took:arg2; their:poss

Concat head token and role



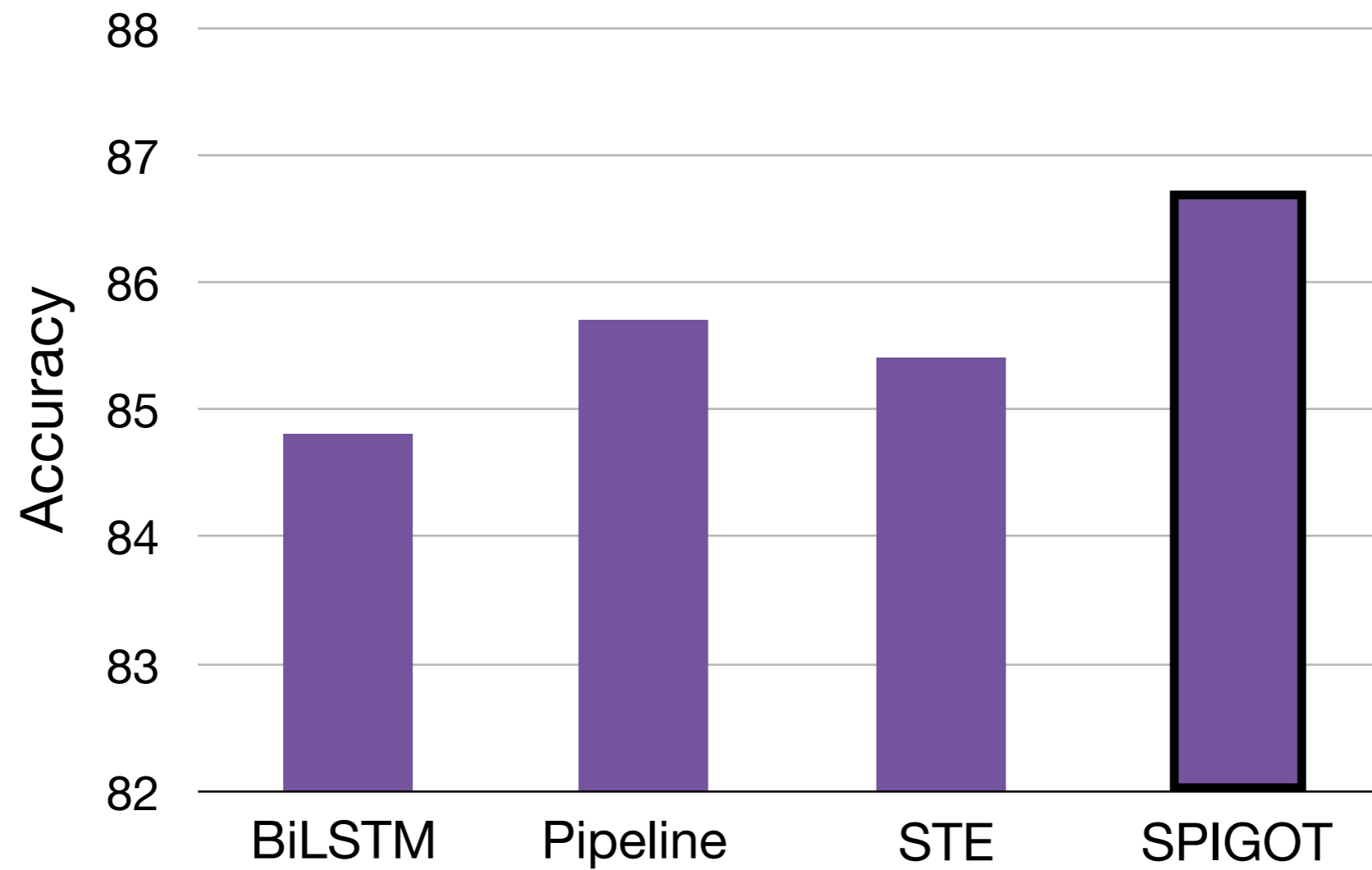
Classifier



BiLSTM+MLP

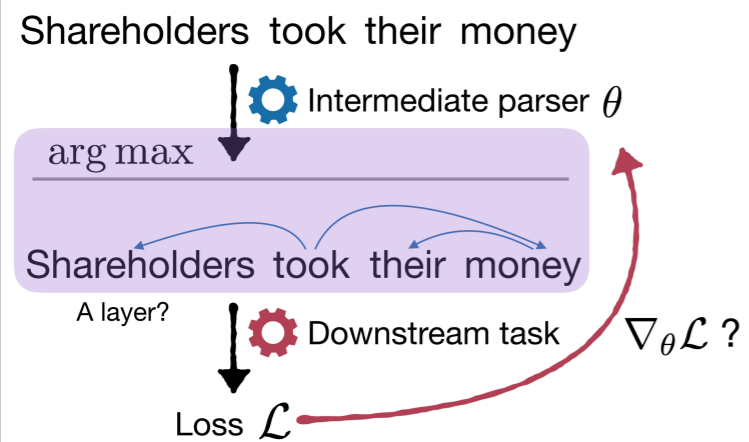
Positive? Negative?

Stanford Sentiment Treebank accuracy



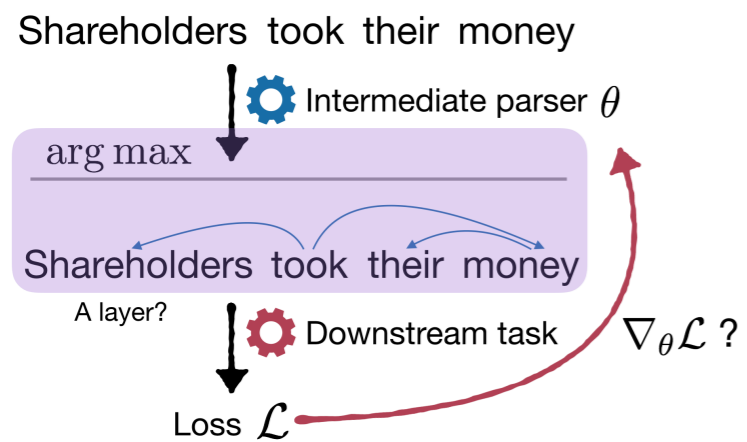
Conclusion

Problem



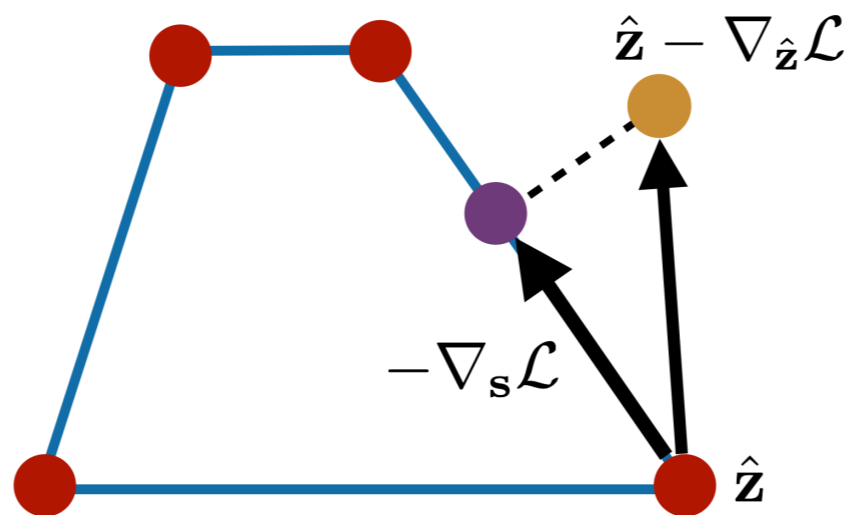
Conclusion

Problem



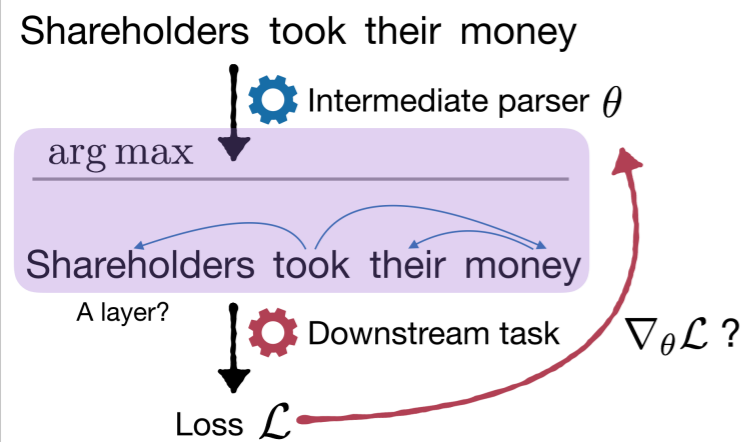
Method

SPIGOT



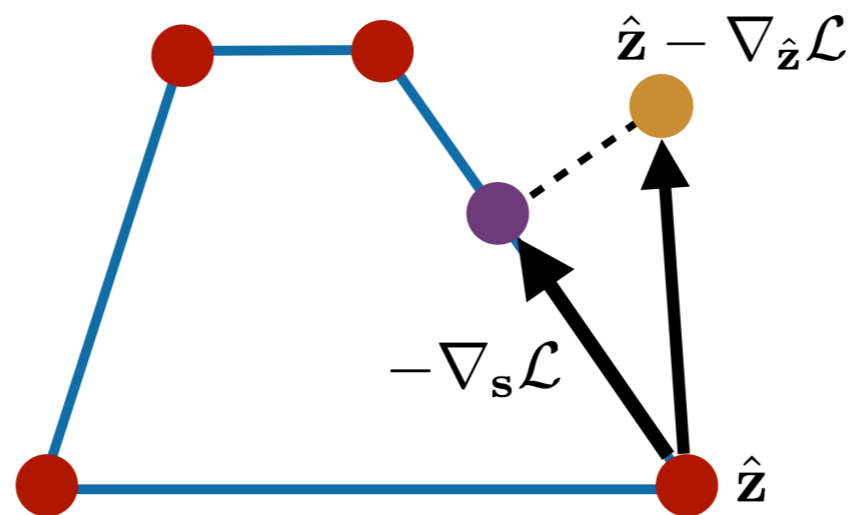
Conclusion

Problem



Method

SPIGOT



Results

SemEval '15. Micro-averaged labeled F_1



Thank you!