Deep Multitask Learning for Semantic Dependency Parsing

ACL17

Hao Peng Sam Thomson. Noah A. Smith

Reporter：Xiachong Feng

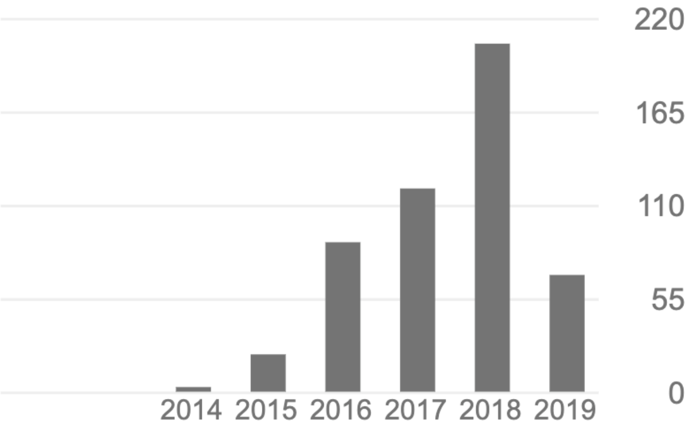
Outline

- Author
- Multitask
- Semantic Dependency Parsing
- Problem
- Motivation
- Model

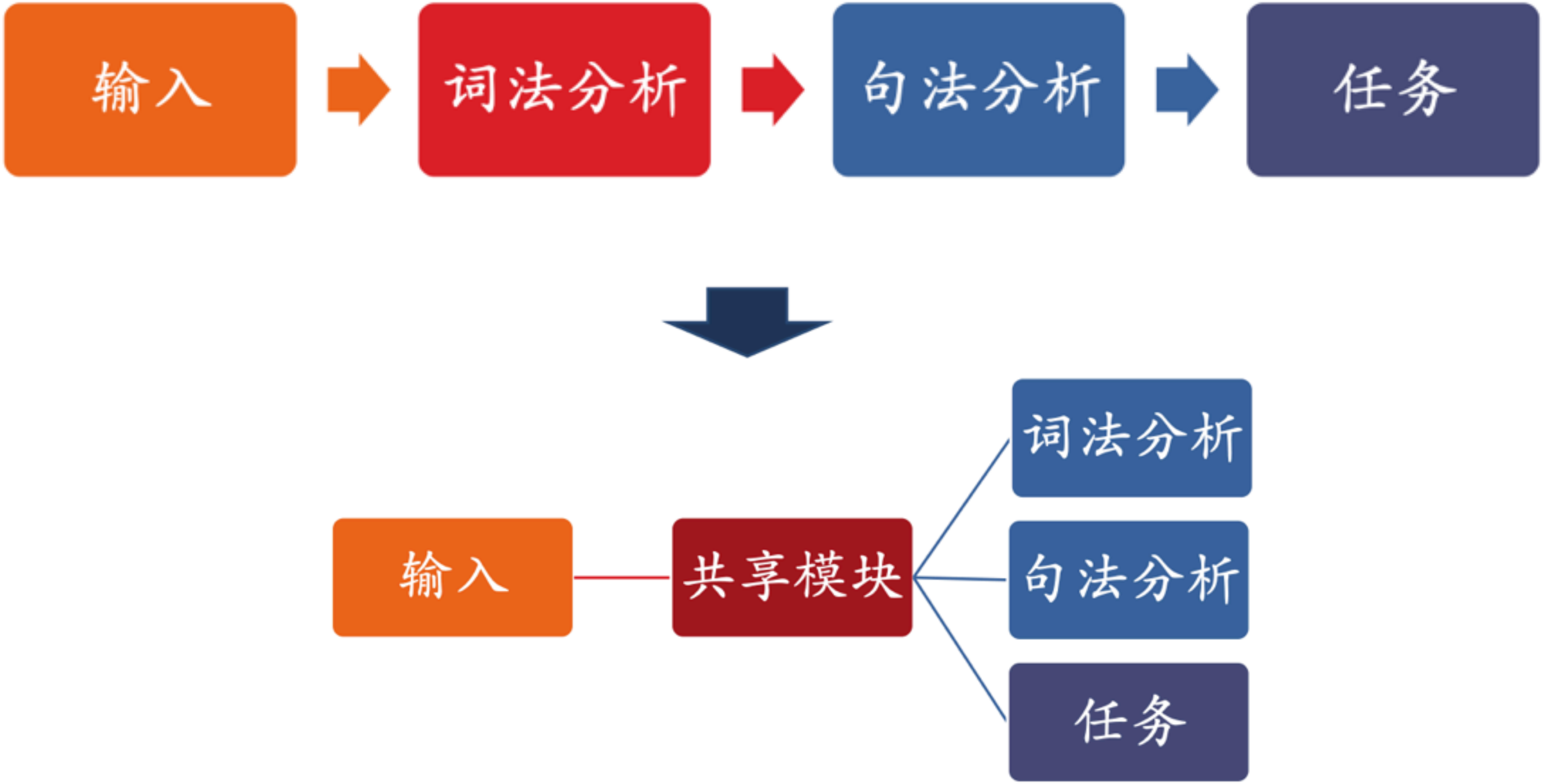
Author

- **Hao Peng**
- Third year Ph.D. student at the University of Washington, advised by Prof. Noah Smith.
- Before coming to UW, he was an undergraduate at **Peking University**.

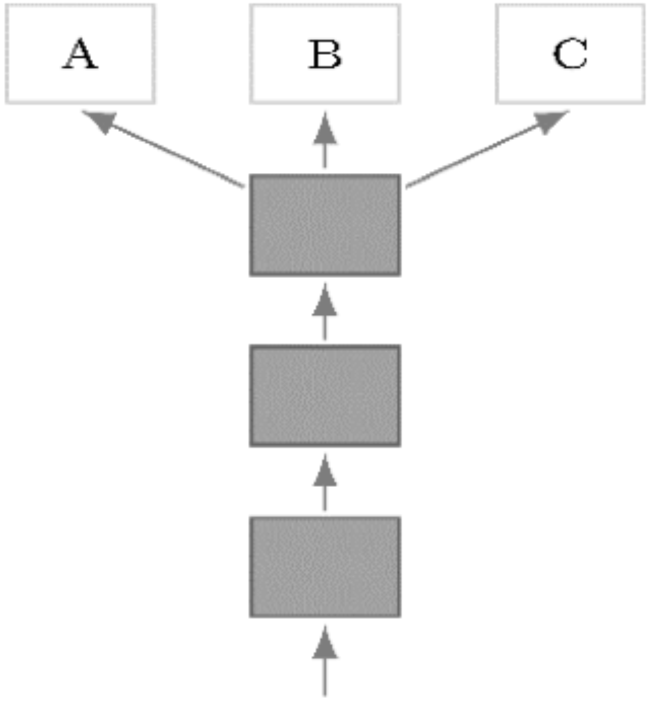
	All	Since 2014
Citations	509	509
h-index	7	7
i10-index	6	6



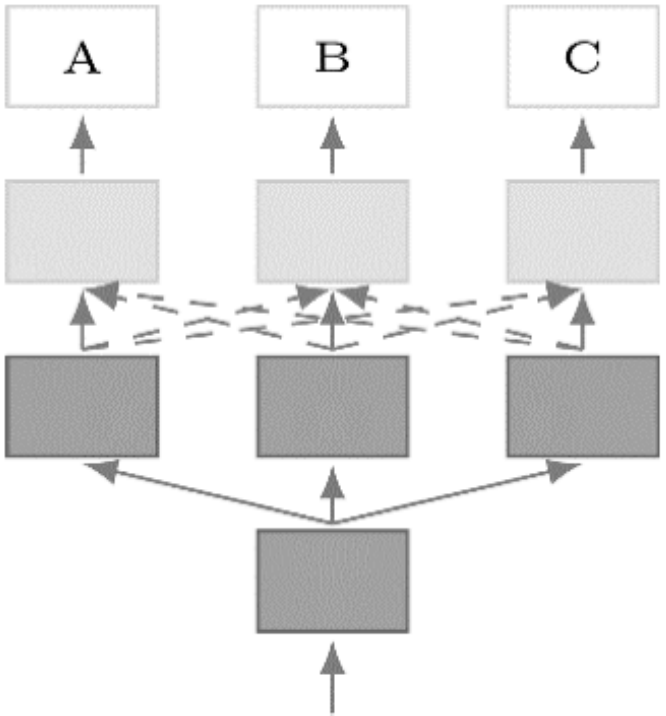
Multitask（多任务学习）



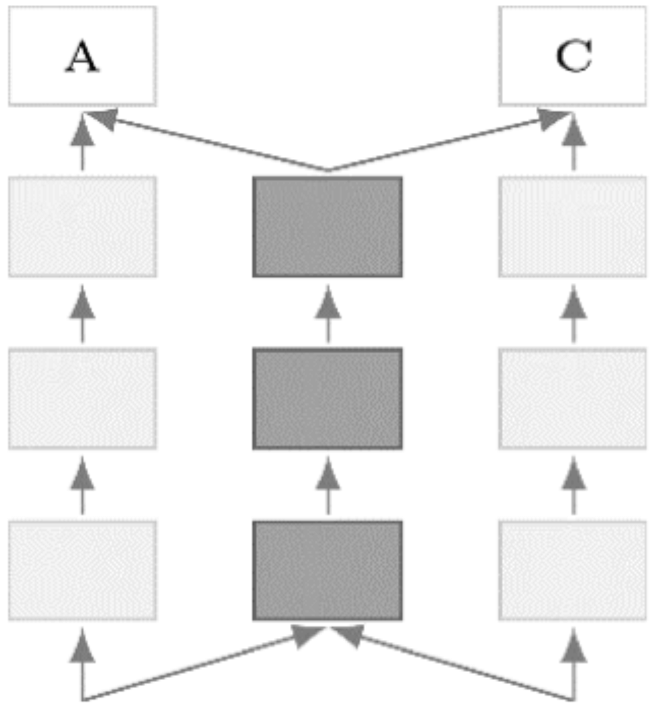
Multitask（多任务学习）



硬共享



软共享



共享私有

自然语言处理中的多任务学习 邱锡鹏

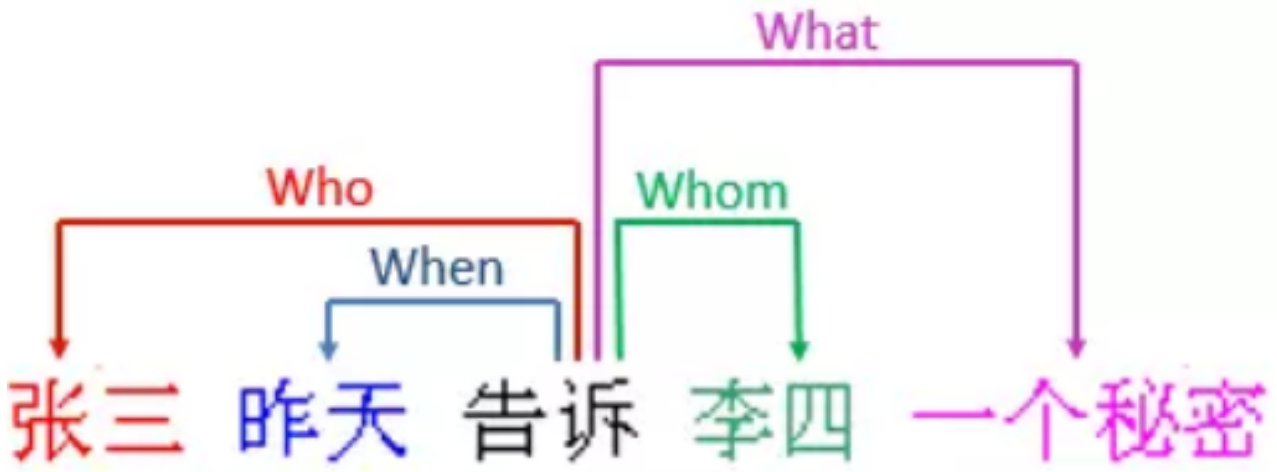
Semantic Dependency Parsing

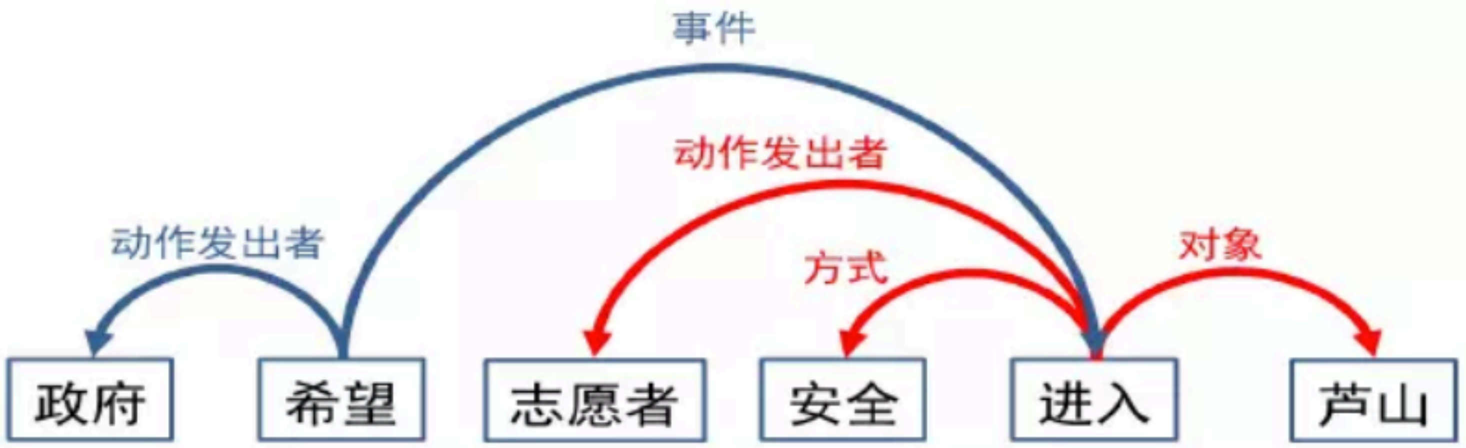
- 语义依存分析：该任务试图找出所有在语义上有所关联的词语对，并且预测相应的语义标签。
- 在中文界，最有影响力的标注方案是BH-SDP，由北京语言大学和哈尔滨工业大学联合制定。

- 语义依存成立的两个词语常常满足：
  - 一个是谓词（predicate），包括大部分谓语性成分（**大部分动词**、小部分名词或形容词）。
  - 另一个是论元（argument），指的是与谓语直接相关的词语（比如谓词是“吃”的话，那么论元就包括“吃”这个动作的发出者、与“吃”相关的食物、餐具、时间和地点等）

- **Who did what to whom when and where**

Who did **what** to **whom when** and **where**

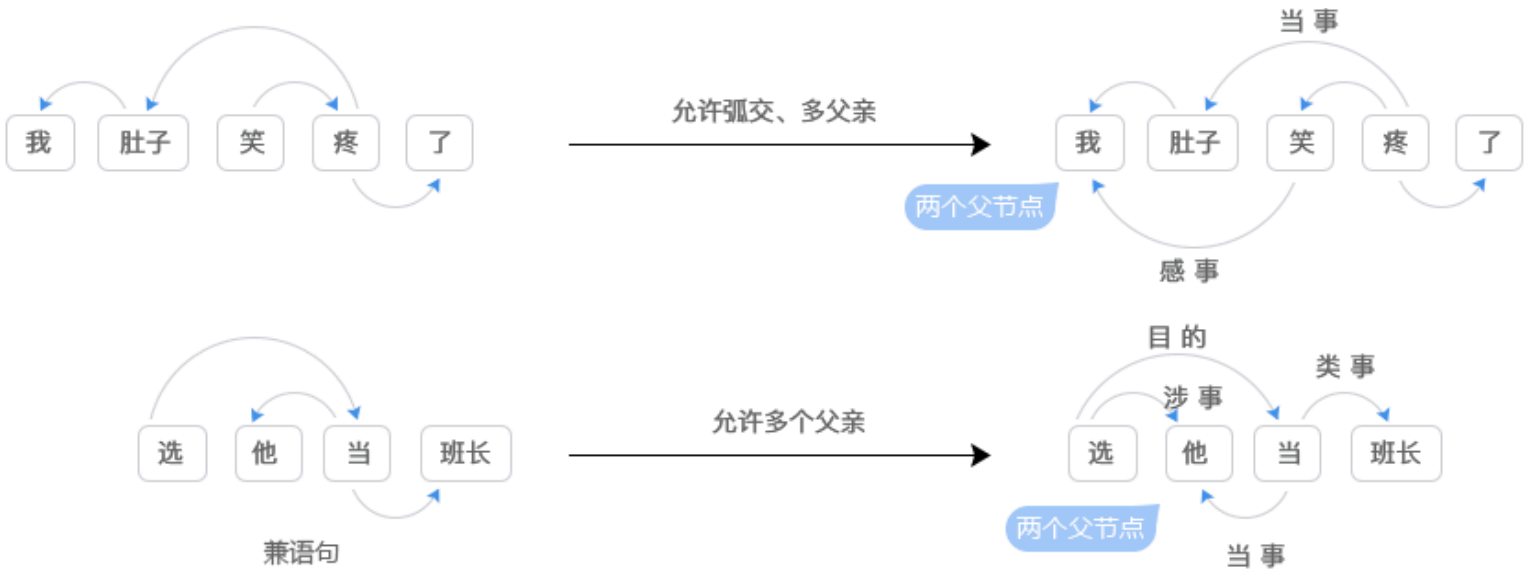




中文语义依存分析—通往中文语义理解的一条蹊径 SCIR丁宇

Semantic Dependency **Tree & Graph**

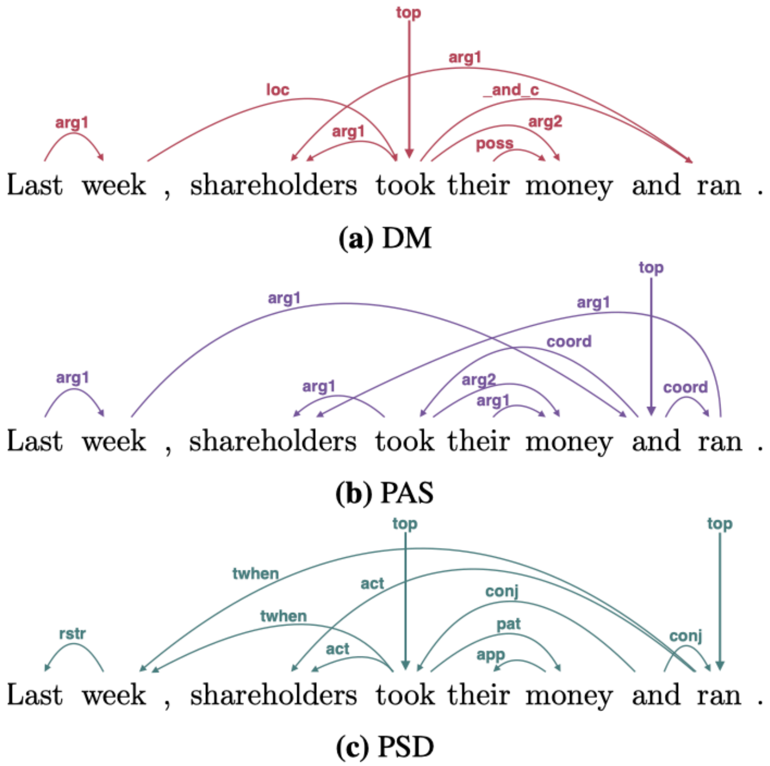
- 语义依存**树** && 语义依存**图**
- 语义依存树与语义依存图的主要区别在于，
    1. 在依存树中，任何一个成分都不能依存于两个或两个以上的成分，而在依存图中则允许句中成分依存于两个或两个以上的成分。
    2. 在依存图中允许依存弧之间存在交叉，而依存树中不允许。



https://www.xfyun.cn/services/semanticDependence

Problem

- Full semantic graphs can be **expensive to annotate**.
- Efforts are fragmented across competing semantic theories, leading to a **limited number of annotations in any one formalism**.

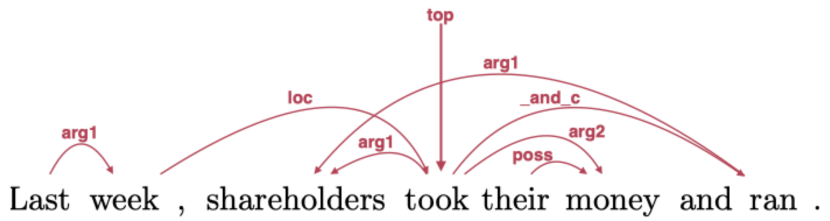


2015 SemEval shared task on Broad-Coverage Semantic Dependency Parsing (SDP; Oepen et al., 2015)

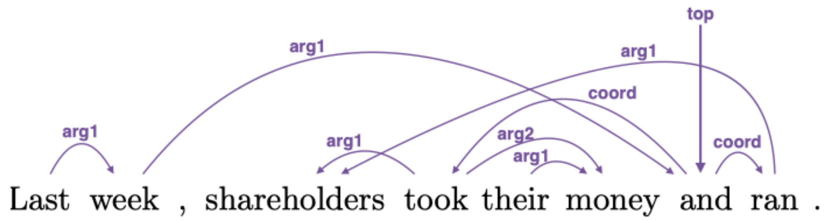
English-language corpus with parallel annotations for **three semantic graph representations**

Motivation

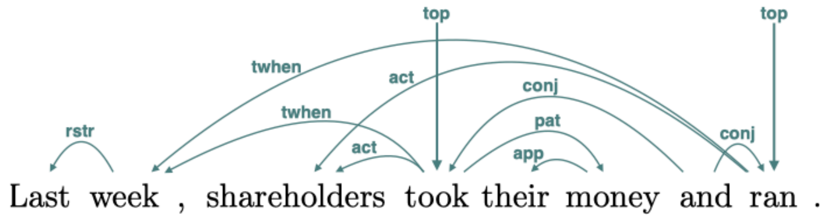
- **Overlap among theories** and **their corresponding representations** can be exploited using multitask learning. allowing us to learn from more data.



(a) DM



(b) PAS

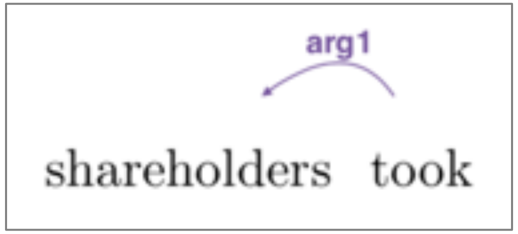


(c) PSD

Three formalisms

- **DM (DELPH-IN MRS)**
  - DeepBank
  - Manually-corrected parses from the LinGO English Resource Grammar
- **PAS (Predicate-Argument Structures)**
  - Extracted from the Enju Treebank
  - Automatic parses from the Enju HPSG parser
- **PSD (Prague Semantic Dependencies)**
  - Extracted from the tectogrammatical layer of the Prague Czech-English Dependency Treebank

Single-Task SDP



- Input sentence x,
- Set of possible semantic graphs $\mathcal{Y}(x)$
- Score function S:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}(x)} S(x, y),$$

- Decompose *S* into a sum of local scores *s* for local structures *p* in the graph

$$S(x, y) = \sum_{p \in y} s(p).$$

- Basic model: Neural arc-factored(弧分解) graph-based dependency parsing
- $\boldsymbol{AD^3}$ to find the highest-scoring internally consistent semantic graph.

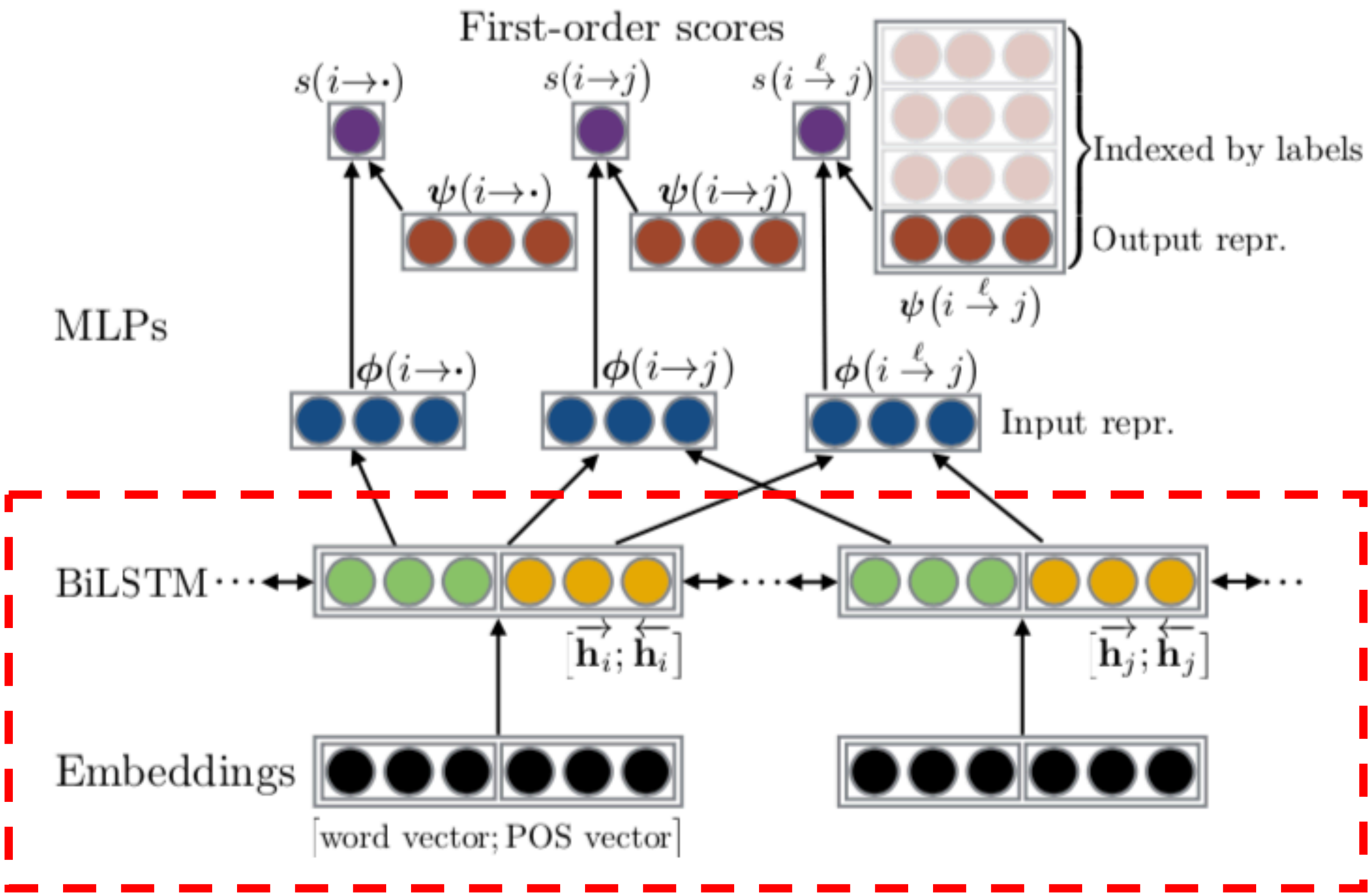
Basic Structure

$\boxed{\text{predicate,}}$ indicating a predicate word, denoted $i \rightarrow \cdot$;

$\boxed{\text{unlabeled arc,}}$ representing the existence of an arc from a predicate to an argument, denoted $i \rightarrow j$;

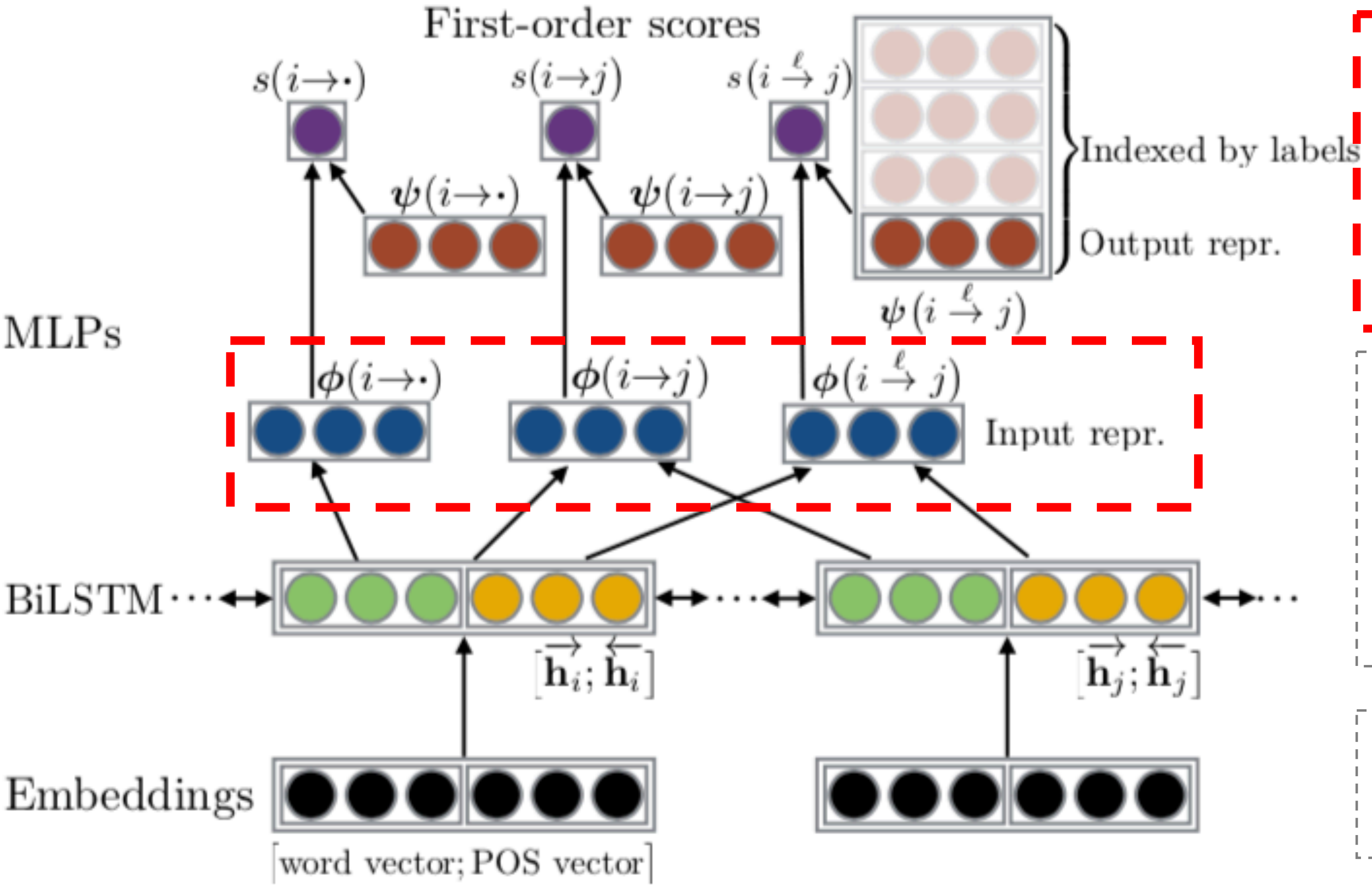
$\boxed{\text{labeled arc,}}$ an arc labeled with a semantic role, denoted $i \xrightarrow{\ell} j$.

Basic Model



First-order scores
$s(i \to \cdot)$
$s(i \to j)$
$s(i \xrightarrow{\ell} j)$
Indexed by labels
Output repr.
$\psi(i \to \cdot)$
$\psi(i \to j)$
$\psi(i \xrightarrow{\ell} j)$
MLPs
$\phi(i \to \cdot)$
$\phi(i \to j)$
$\phi(i \xrightarrow{\ell} j)$
Input repr.
BiLSTM
$[\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$
$[\overrightarrow{\mathbf{h}}_j; \overleftarrow{\mathbf{h}}_j]$
Embeddings
[word vector; POS vector]

Basic Model

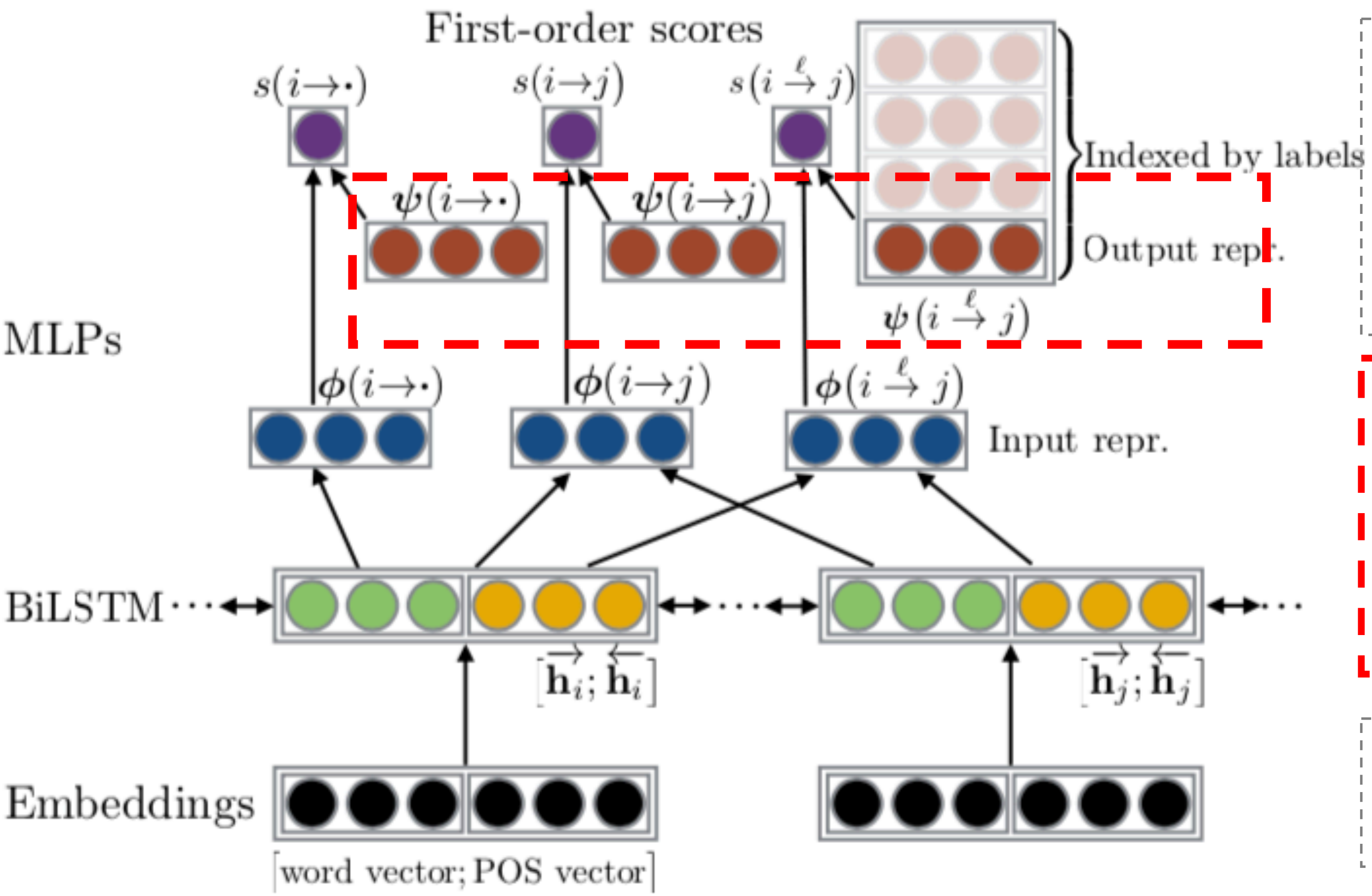


$$\phi(i\to\cdot) = \tanh\big(\mathbf{C}_{\text{pred}}\mathbf{h}_i + \mathbf{b}_{\text{pred}}\big)$$
$$\phi(i\to j) = \tanh\big(\mathbf{C}_{\text{UA}}[\mathbf{h}_i;\mathbf{h}_j] + \mathbf{b}_{\text{UA}}\big),$$
$$\phi(i \xrightarrow{\ell} j) = \tanh\big(\mathbf{C}_{\text{LA}}[\mathbf{h}_i;\mathbf{h}_j] + \mathbf{b}_{\text{LA}}\big).$$

$$\boldsymbol{\psi}(i\to\cdot) = \boldsymbol{\psi}_{\text{pred}},$$
$$\boldsymbol{\psi}(i\to j) = \boldsymbol{\psi}_{\text{UA}},$$
$$\boldsymbol{\psi}(i \xrightarrow{\ell} j) = \boldsymbol{\psi}_{\text{LA}}(\ell).$$

$$s(p) = \boldsymbol{\phi}(p)\cdot\boldsymbol{\psi}(p).$$

Basic Model



$$\phi(i\to\cdot) = \tanh\left(\mathbf{C}_{\text{pred}}\mathbf{h}_i + \mathbf{b}_{\text{pred}}\right)$$

$$\phi(i\to j) = \tanh\left(\mathbf{C}_{\text{UA}}\left[\mathbf{h}_i; \mathbf{h}_j\right] + \mathbf{b}_{\text{UA}}\right),$$

$$\phi(i \xrightarrow{\ell} j) = \tanh\left(\mathbf{C}_{\text{LA}}\left[\mathbf{h}_i; \mathbf{h}_j\right] + \mathbf{b}_{\text{LA}}\right).$$

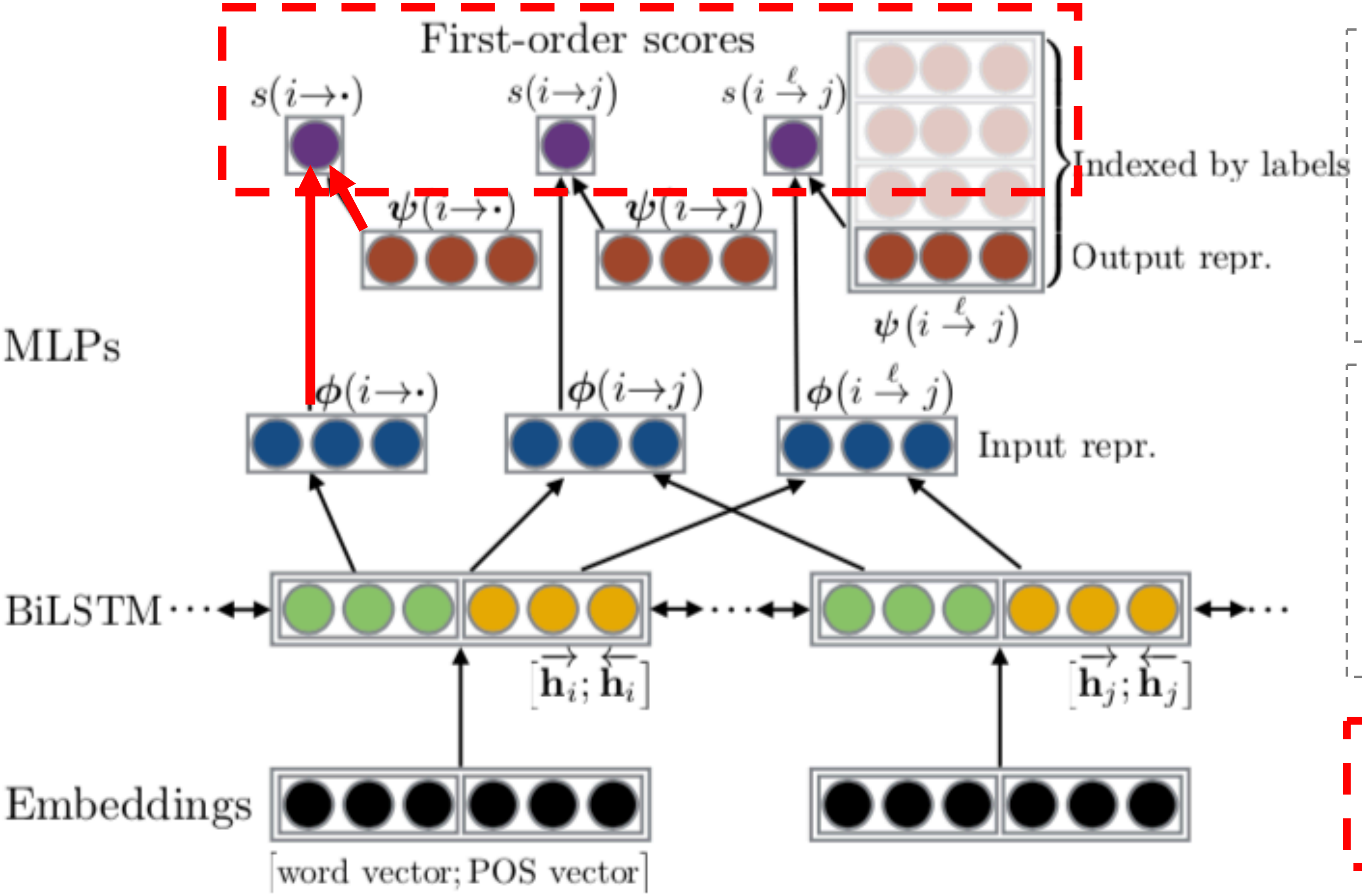
$$\psi(i\to\cdot) = \psi_{\text{pred}},$$

$$\psi(i\to j) = \psi_{\text{UA}},$$

$$\psi(i \xrightarrow{\ell} j) = \psi_{\text{LA}}(\ell).$$

$$s(p) = \phi(p) \cdot \psi(p).$$

Basic Model



$$\phi(i\rightarrow\cdot) = \tanh\left(\mathbf{C}_{\text{pred}}\mathbf{h}_i + \mathbf{b}_{\text{pred}}\right)$$

$$\phi(i\rightarrow j) = \tanh\left(\mathbf{C}_{\text{UA}}\left[\mathbf{h}_i;\mathbf{h}_j\right] + \mathbf{b}_{\text{UA}}\right),$$

$$\phi(i\xrightarrow{\ell}j) = \tanh\left(\mathbf{C}_{\text{LA}}\left[\mathbf{h}_i;\mathbf{h}_j\right] + \mathbf{b}_{\text{LA}}\right).$$

$$\psi(i\rightarrow\cdot) = \psi_{\text{pred}},$$

$$\psi(i\rightarrow j) = \psi_{\text{UA}},$$

$$\psi(i\xrightarrow{\ell}j) = \psi_{\text{LA}}(\ell).$$

$$s(p) = \phi(p)\cdot\psi(p).$$

Learning

- Loss function

$$\min_{\Theta} \frac{\lambda}{2}\|\Theta\|^2 + \frac{1}{N}\sum_{i=1}^{N} L(x_i, y_i; \Theta),$$

L2-regularized

structured hinge loss

$$L(x_i, y_i; \Theta) = \max_{y \in \mathcal{Y}(x_i)} \{S(x_i, y) + c(y, y_i)\} - S(x_i, y_i).$$

Sentence Gold parse

Decoding Constraints

$i \rightarrow \cdot$ if and only if there exists at least one j such that $i \rightarrow j$;

If $i \rightarrow j$, then there must be exactly one label ℓ such that $i \xrightarrow{\ell} j$. Conversely, if not $i \rightarrow j$, then there must not exist any $i \xrightarrow{\ell} j$;

Experiments

	Model	DM	PAS	PSD	Avg.
id	Du et al., 2015	89.1	91.3	75.7	86.3
	A&M, 2015	88.2	90.9	76.4	86.0
	BASIC	**89.4**	**92.2**	**77.6**	**87.4**
ood	Du et al., 2015	81.8	87.2	73.3	81.7
	A&M, 2015	81.8	86.9	74.8	82.0
	BASIC	**84.5**	**88.3**	**75.3**	**83.6**

Table 2: Labeled parsing performance (F_1 score) on both in-domain (id) and out-of-domain (ood) test data. The last column shows the micro-average over the three tasks. Bold font indicates best performance without syntax. Underlines indicate statistical significance with Bonferroni (1936) correction compared to the best baseline system.[4]

Multitask SDP

- Use training data for all three formalisms to improve performance on each formalism's parsing task.

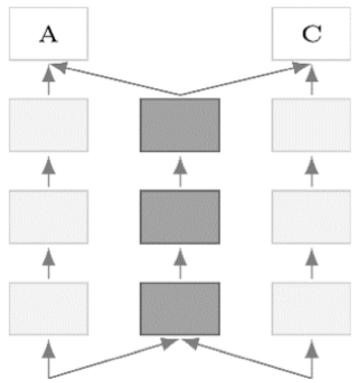
- **First-order model**, where representation functions are enhanced by parameter sharing while inference is kept separate for each task

- **Cross-task higher-order structures** that uses joint inference *across* different tasks

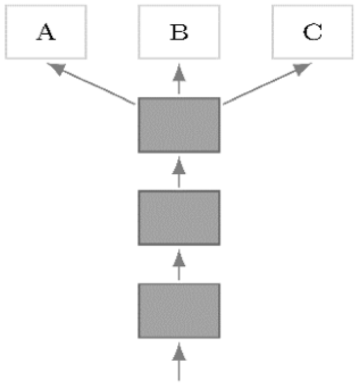
Multitask SDP with Parameter Sharing

- **FREDA** :Task-specific BiLSTM encoders as well as a **common one that is shared across all tasks($\widetilde{h}$).**

$$\phi^{(t)}(i \xrightarrow{\ell} j) = \tanh\big(\mathbf{C}^{(t)}_{\text{LA}}[\mathbf{h}^{(t)}_i; \mathbf{h}^{(t)}_j; \boxed{\widetilde{\mathbf{h}}_i; \widetilde{\mathbf{h}}_j]} + \mathbf{b}^{(t)}_{\text{LA}}\big).$$

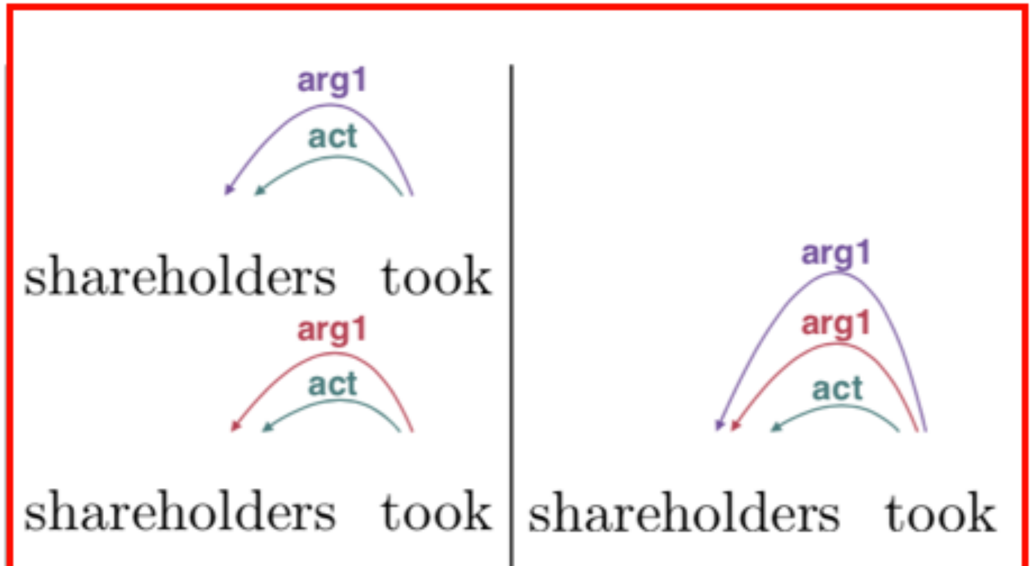


- **SHARED**: use only the shared encoder and does not use task-specific encoders



Multitask SDP with Cross-Task Structures

- Look at interactions between **arcs** that share the **same head** and **modifier**



(b) Second-order. (c) Third-order.

Multitask SDP with Cross-Task Structures

- Higher-order structure scoring

$$\sum_{j=1}^{r} \prod_{t \in \mathcal{T}} \left[\underbrace{\mathbf{U}_{\mathrm{LA}}^{(t)}}_{\text{parameter}} \boldsymbol{\phi}^{(t)}\left(p^{(t)}\right)\right]_j \left[\underbrace{\mathbf{V}_{\mathrm{LA}}^{(t)}}_{\text{parameter}} \boldsymbol{\psi}^{(t)}\left(p^{(t)}\right)\right]_j .$$

Experiments

- **SHARED1**
  - First-order model
  - Single shared Bi-LSTM encoder
  - Inference separate for each task
- **FREDA1**
  - First-order model
  - Shared encoder as well as task-specific ones
  - Inference is kept separate for each task
- **SHARED3**
  - Third-order model
  - Shared Bi-LSTM encoder
  - Cross-task structures and inference
- **FREDA3**
  - Third-order model
  - Shared encoder as well as task-specific ones
  - Cross-task structures and inference

Experiments

	DM	PAS	PSD	Avg.
Du et al., 2015	89.1	91.3	75.7	86.3
A&M, 2015 (closed)	88.2	90.9	76.4	86.0
A&M, 2015 (open)†	89.4	91.7	77.6	87.1
BASIC	89.4	92.2	77.6	87.4
SHARED1	89.7	91.9	77.8	87.4
FREDA1	90.0	92.3	78.1	87.7
SHARED3	90.3	92.5	**78.5**	**88.0**
FREDA3	**90.4**	**92.7**	**78.5**	**88.0**

(a) Labeled F_1 score on the in-domain test set.

- Even with the best open track system for **DM and PSD**, but improves on **PAS and on average**, without making use of any syntax.

Experiments

	DM	PAS	PSD	Avg.
Du et al., 2015	89.1	91.3	75.7	86.3
A&M, 2015 (closed)	88.2	90.9	76.4	86.0
A&M, 2015 (open)†	89.4	91.7	77.6	87.1
BASIC	89.4	92.2	77.6	87.4
SHARED1	89.7	91.9	77.8	87.4
FREDA1	90.0	92.3	78.1	87.7
SHARED3	90.3	92.5	**78.5**	**88.0**
FREDA3	**90.4**	**92.7**	**78.5**	**88.0**

(a) Labeled F_1 score on the in-domain test set.

- Even with the best open track system for DM and PSD, but improves on PAS and on average, without making use of any syntax.
- **Three of our four** multitask variants further improve over our basic model .

Experiments

	DM	PAS	PSD	Avg.
Du et al., 2015	89.1	91.3	75.7	86.3
A&M, 2015 (closed)	88.2	90.9	76.4	86.0
A&M, 2015 (open)†	89.4	91.7	77.6	87.1
BASIC	89.4	92.2	77.6	87.4
SHARED1	89.7	91.9	77.8	87.4
FREDA1	90.0	92.3	78.1	87.7
SHARED3	90.3	92.5	**78.5**	**88.0**
FREDA3	**90.4**	**92.7**	**78.5**	**88.0**

(a) Labeled F_1 score on the in-domain test set.

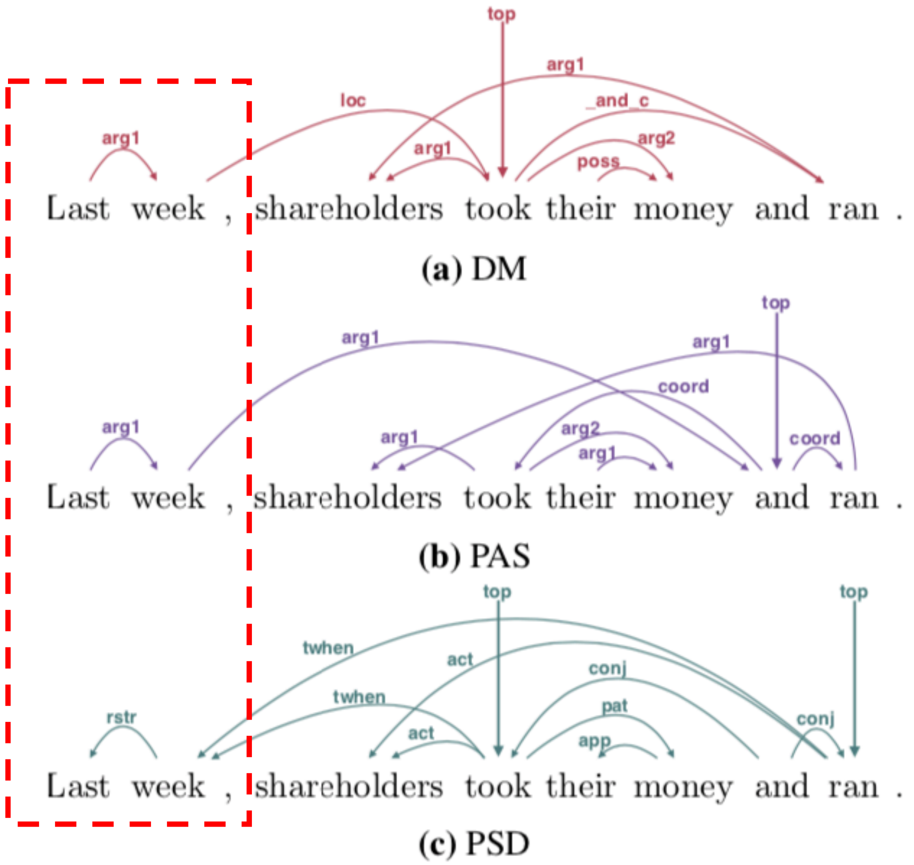
- Even with the best open track system for DM and PSD, but improves on PAS and on average, without making use of any syntax.
- Three of our four multitask variants further improve over our basic model .
- **Best** models (SHARED3, FREDA3)

Experiments-Effects of structural overlap

- **DM and PAS** are more structurally similar to each other than either is to PSD.

	Undirected			Directed		
	DM	PAS	PSD	DM	PAS	PSD
DM	-	67.2	56.8	-	64.2	26.1
PAS	70.0	-	54.9	66.9	-	26.1
PSD	57.4	56.3	-	26.4	29.6	-

Table 5: Pairwise structural similarities between the three formalisms in unlabeled F_1 score. Scores from Oepen et al. (2015).



Experiments-Effects of structural overlap

- improves on DM and PAS, but *degrades* on PSD.

	DM		PAS		PSD	
	UF	LF	UF	LF	UF	LF
FREDA1	91.7	90.4	93.1	91.6	89.0	79.8
FREDA3	91.9	90.8	93.4	92.0	88.6	80.4

Table 6: Unlabeled (UF) and labeled (LF) parsing performance of FREDA1 and FREDA3 on the development set of SemEval 2015 Task 18.

Thanks!