Exploration on Few-shot Slot Tagging

Speaker : Yutai Hou **Mentor:** Wanxiang Che

Research Center for Social Computing and Information Retrieval - Harbin Institute of Technology

Homepage : https://atmahou.github.io/

Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, Ting Liu

Introduction to Few-shot Learning

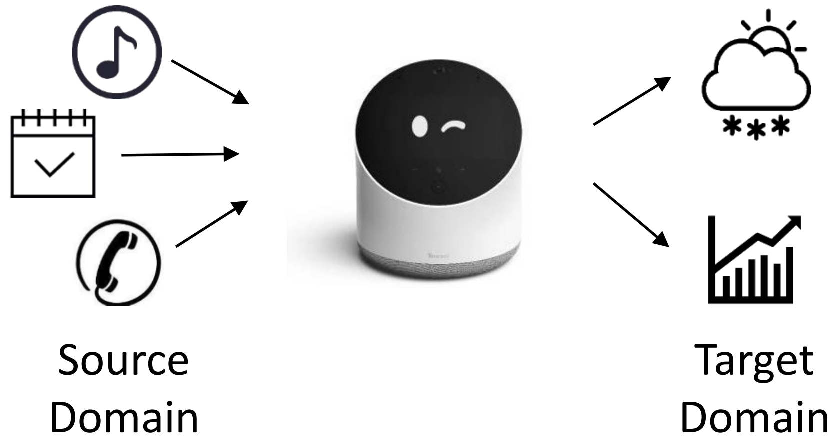
- Background
  - Deep learning methods have achieved great success, but require lots of labeled data
  - Human are good at learning new knowledge through a few instances:
    - For example, a child only needs one or two pictures to recognize lions and zebras

Humans don’t learn from scratch.
Experience accumulated from birth assists learning.

  - We hope that model can use prior experience to learn from a few examples.
    - Make artificial intelligence less artificial, more intelligent

Introduction to Few-shot Learning

- Background
  - Real-world applications often face **new demands** and **new domains**:
    - New domains often lack data.
    - Annotation schema needs adjustment: add, del labels
    - In industrial scenarios, the cost of frequent retraining and deployment of models is often unacceptable

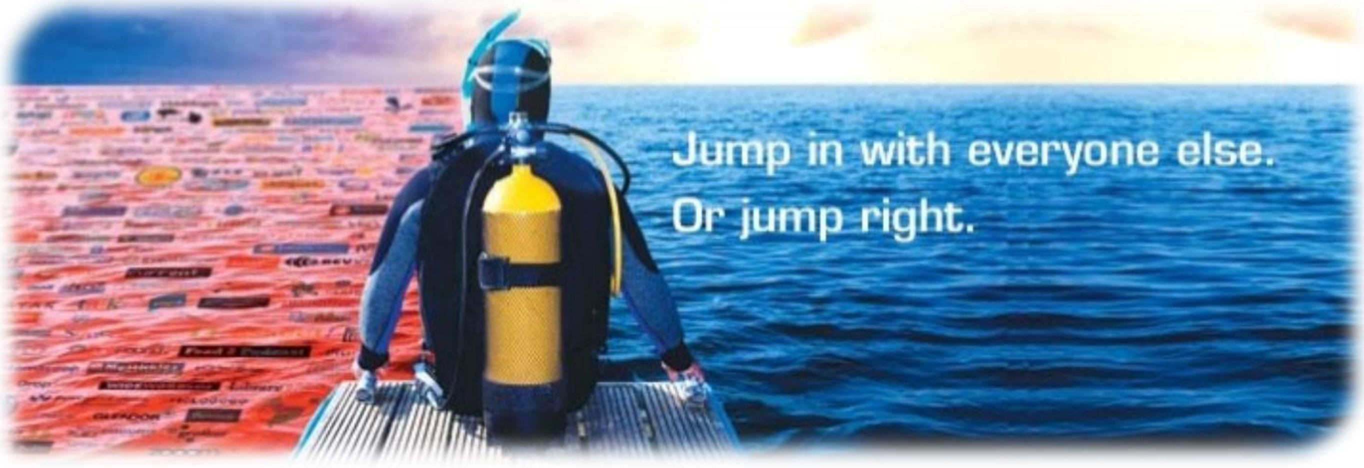


  - Many companies are working on data scarcity issues:

Introduction to Few-shot Learning

- Few-shot Learning
  - It is a technology that specifically solves the aforementioned problems.
  - In the general context, *Few-shot Learning* refers to an application of *Meta Learning*

- Few-shot learning for NLP is still less investigated:
  - especially for specific NLP application, such as **dialogue**.
  - especially for structure prediction problem, such as **sequence labeling**.



Introduction to Slot Tagging

- Task-oriented Dialogue
  - A dialogue system to help users complete specific goals, such as hotel and flight reservations.

- Slot Tagging
  - a key module in the task-oriented dialogue
  - is usually formulated as a sequence labeling problem
  - Example：

User Utterance

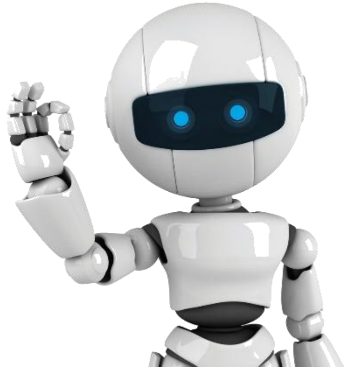
show me the closest restaurant.

↓

Slot Tagging

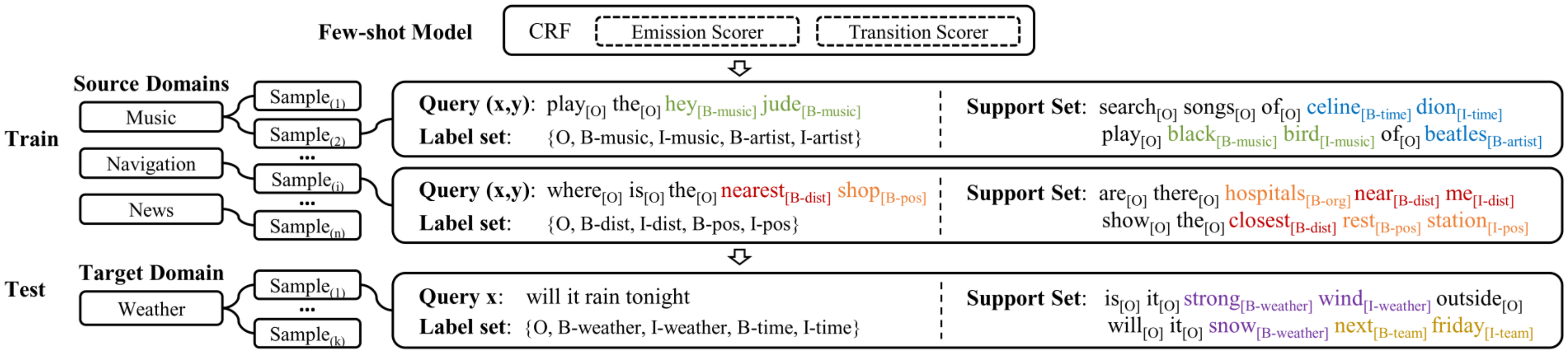
↓

show me the [closest]<distance> [restaurant]<poi type>



Few-shot Slot-Tagging

- Train：
  - Train on data-rich domains
  - To keep training consistent with the test, simulate the few-shot setting during training (Vinyals et al., 2016)
- Test：
  - Test on unseen new domains
  - Given a few-shot support set，tagging slots for unseen query sentences.



Overviews of training and testing for Few-shot Slot Tagging

Challenge for Few-shot Slot Tagging

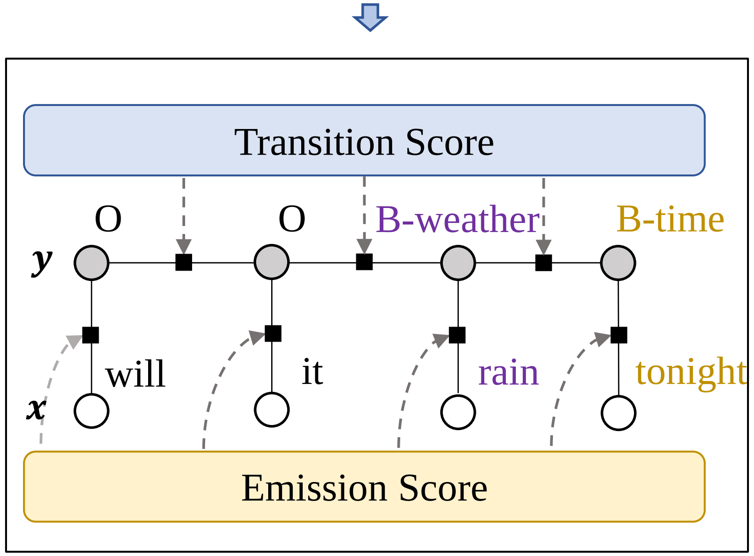
Transition:

—— Requires to consider dependency between labels.

- Due to the discrepancy of label sets, previously learned label dependencies are not directly transferable.
- It is hard to learn reasonable label dependencies from only a few examples.

Query sentence

***x*: will it rain tonight**



Traditional Slot-Tagging Framework

Emissions:

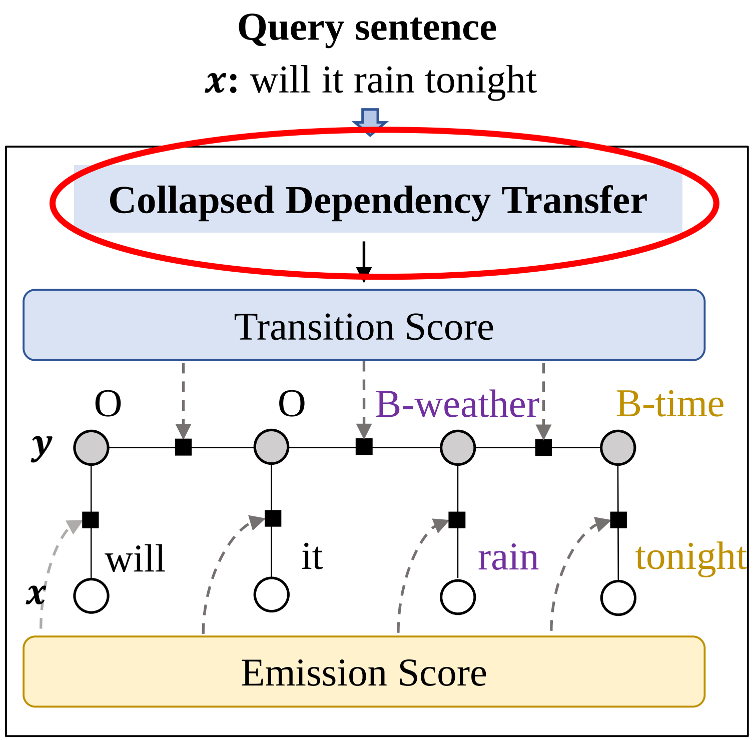
—— Computed as token-label similarity in few-shot setting.

- Label representations often distribute closely in the embedding space and cause misclassification.
- A word tends to mean differently in different context and domain.

Challenges and Solutions for Transition Score

- **Problem:**
  - Due to the discrepancy of label sets, previously learned label dependencies are not directly transferable.
  - It is hard to learn reasonable label dependencies from only a few examples.

- **Idea:**
  - Learn and transfer abstract label dependency patterns.
  - In target domains, expand the abstract label dependency into domain-specific dependencies.



Proposed CRF Framework

Challenges and Solutions for Transition Score

- Solution: **Collapsed Dependency Transfer**
  - Learn a collapsed transition matrix to model abstract label dependencies.
  - In target domain, expand it to fit target label set.

	O	sB	dB	sI	dI
O	0.6	0.4	\	0	\
B	0.3	0.1	0.1	0.5	0
I	0.6	0.1	0.1	0.2	0

Collapsed Label Transition $\tilde{\mathbf{T}}$

Expand

	O	B-time	B-weather	I-time	I-weather
O	0.6	0.4	0.4	0	0
B-time	0.3	0.1	0.1	0.5	0
B-weather	0.3	0.1	0.1	0	0.5
I-time	0.6	0.1	0.1	0.2	0
I-weather	0.6	0.1	0.1	0	0.2

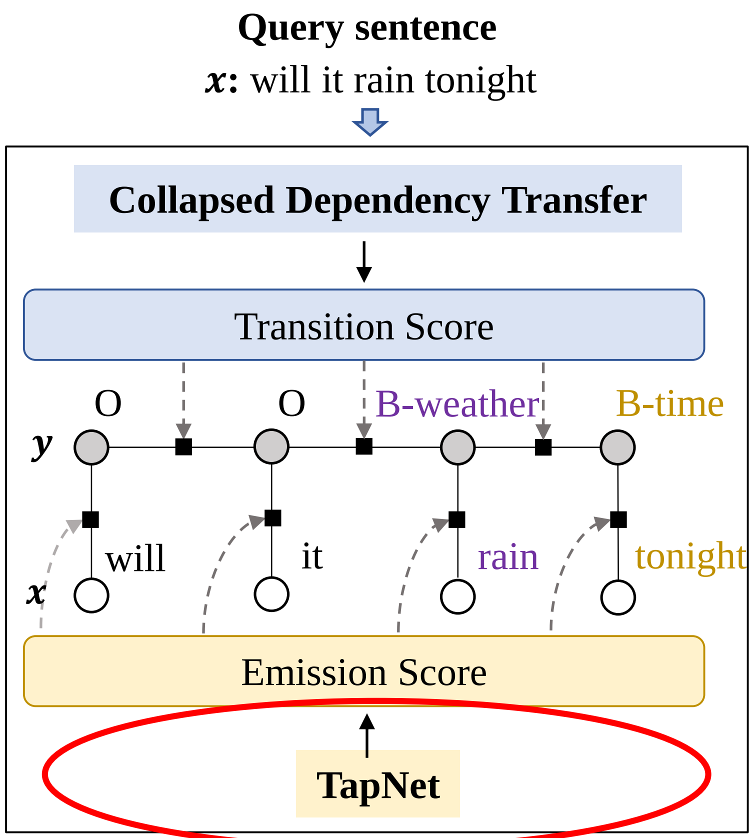
Expanded Label Transition $\mathbf{T}$

Challenges and Solutions for Emission Score

- **Problem:**
  - Label representations often distribute closely in the embedding space and cause misclassification.

- **Idea:**
  - Pull away label representations to make them well-separated
  - Exploit semantic info within label name strings

- **Solution: Task-adaptive Projection Net (TapNet)**
  - Find task-adaptive project spaces where different categories are well-separated from each other.
  - Use *Linear Error Nulling* to construct such projection space (Yoon et al. ICML 2019)



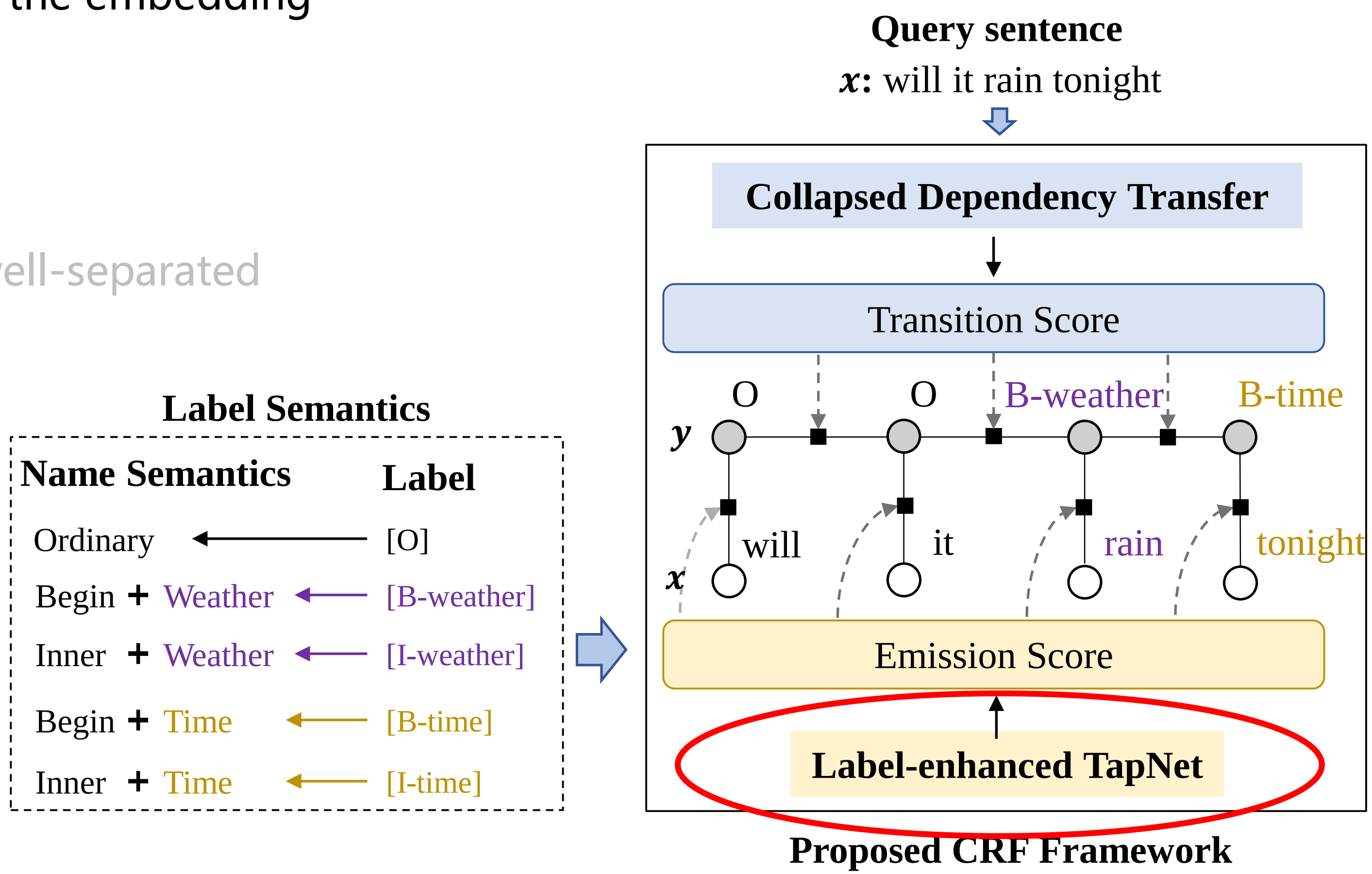
Proposed CRF Framework

Challenges and Solutions for Emission Score

- **Problem:**
  - Label representations often distribute closely in the embedding space and cause misclassification.

- **Idea:**
  - Pull away label representations to make them well-separated
  - Exploit semantic info within label name strings

- **Solution: Label-enhanced TapNet**
  - When constructing the projection space, require to align the label name with prototypes
  - Represent label with both label name embedding and prototypes

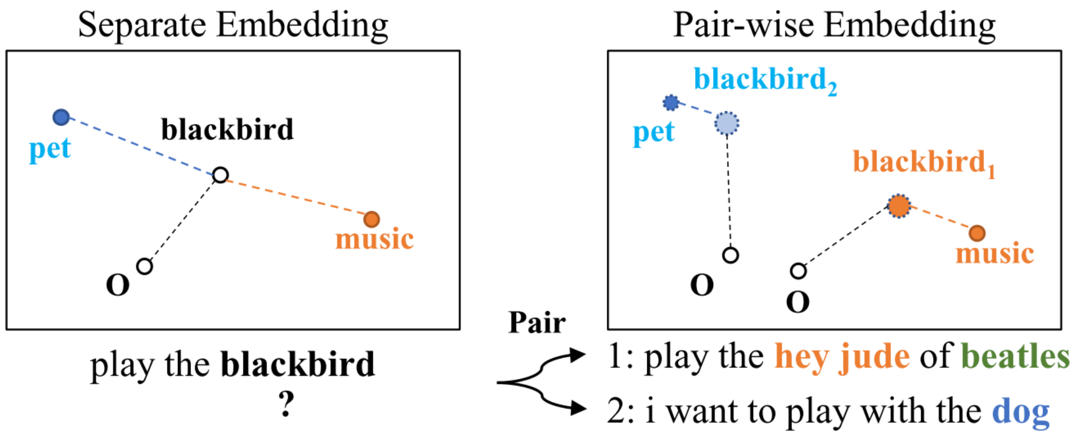


Challenges and Solutions for Emission Score

- Problem:
  - A word tends to mean differently in different context and domain.

- Idea:
  - Use context to disambiguate words.
  - Use domain-specific context provided by sentences in support set

- Solution: Pair-wise Embedding
  - Embed query and support words pair-wisely.
  - Capture context with BERT



Experiment Data

- We conducted experiments on 5 datasets from two tasks:

Task	Dataset	Domain	# Sent	# Labels
Slot Tagging	Snips	Weather	2,100	10
		Music	2,100	10
		PlayList	2,042	6
		Book	2056	8
		SearchScreen	2,059	8
		Restaurant	2,073	15
		CreativeWork	2,054	3
NER	CoNLL	News	20679	5
	GUM	WiKi	3,493	12
	WNUT	Social	5,657	7
	OntoNotes	Mixed	159,615	19

Statistic of Raw Data

Experiment Data

- We construct the few-shot data from the original data
  - Construct 1-shot and 5-shot data
  - Perform cross-evaluation and take the averaged results
    - Each time: 1 test domain, 1 dev domain, rest train domain

Domain	Slot Tagging							Named Entity Recognition			
	We	Mu	Pl	Bo	Se	Re	Cr	News	Wiki	Social	Mixed
Ave. $\|S\|$ (1-shot)	6.15	7.66	2.96	4.34	4.29	9.41	1.30	3.38	6.50	5.48	14.38
Samples (1-shot)	2,000	2,000	2,000	2,000	2,000	2,000	2,000	4,000	4,000	4,000	4,000
Ave. $\|S\|$ (5-shot)	28.91	34.43	13.84	19.83	19.27	41.58	5.28	15.58	27.81	28.66	62.28
Samples (5-shot)	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Statistic of Few-shot Data

Main Result for Slot Tagging

Model	1-shot Slot Tagging								5-shots Slot Tagging							
	We	Mu	Pl	Bo	Se	Re	Cr	Ave.	We	Mu	Pl	Bo	Se	Re	Cr	Ave.
Bi-LSMT	10.36	17.13	17.52	53.84	18.44	22.56	8.64	21.21	25.17	39.80	46.13	74.60	53.47	40.35	25.10	43.52
SimBERT	36.10	37.08	35.11	68.09	41.61	42.82	23.91	40.67	53.46	54.13	42.81	75.54	57.10	55.30	32.38	52.96
TransferBERT	55.82	38.01	45.65	31.63	21.96	41.79	38.53	39.06	59.41	42.00	46.07	20.74	28.20	67.75	58.61	46.11
MN	21.74	10.68	39.71	58.15	24.21	32.88	**69.66**	36.72	36.67	33.67	52.60	69.09	38.42	33.28	72.10	47.98
WPZ	4.53	7.43	14.43	39.15	11.69	7.78	10.09	13.59	9.54	14.23	18.12	44.65	18.98	12.03	14.05	18.80
WPZ+GloVe	17.92	22.37	19.90	42.61	22.30	22.79	16.75	23.52	26.61	34.25	22.11	50.55	28.53	34.16	23.69	31.41
WPZ+BERT	46.72	40.07	50.78	68.73	60.81	55.58	67.67	55.77	67.82	55.99	46.02	72.17	73.59	60.18	66.89	63.24
TapNet	51.12	40.65	48.41	77.50	49.77	54.79	61.39	54.80	53.03	49.80	54.90	83.36	63.07	59.84	67.02	61.57
TapNet+CDT	66.30	55.93	57.55	83.32	64.45	65.65	67.91	65.87	66.48	66.36	68.23	**85.76**	73.60	64.20	68.47	70.44
L-WPZ+CDT	71.23	47.38	59.57	81.98	69.83	66.52	62.84	65.62	**74.68**	56.73	52.20	78.79	80.61	69.59	67.46	68.58
L-TapNet+CDT	**71.53**	**60.56**	**66.27**	**84.54**	**76.27**	**70.79**	62.89	**70.41**	71.64	**67.16**	**75.88**	84.38	**82.58**	**70.05**	**73.41**	**75.01**

Main Result for NER

Model	1-shot Named Entity Recognition					5-shots Named Entity Recognition				
	News	Wiki	Social	Mixed	Ave.	News	Wiki	Social	Mixed	Ave.
Bi-LSMT	2.57 ±0.14	3.29 ±0.19	0.67 ±0.07	2.11 ±0.15	2.16 ±0.14	6.81 ±0.40	8.40 ±0.16	1.06 ±0.16	13.17 ±0.17	7.36 ±0.22
SimBERT	19.22 ±0.00	6.91 ±0.00	5.18 ±0.00	13.99 ±0.00	11.32 ±0.00	32.01 ±0.00	10.63 ±0.00	8.20 ±0.00	21.14 ±0.00	18.00 ±0.00
TransferBERT	4.75 ±1.42	0.57 ±0.32	2.71 ±0.72	3.46 ±0.54	2.87 ±0.75	15.36 ±2.81	3.62 ±0.57	11.08 ±0.57	**35.49** ±7.60	16.39 ±2.89
MN	19.50 ±0.35	4.73 ±0.16	17.23 ±2.75	15.06 ±1.61	14.13 ±1.22	19.85 ±0.74	5.58 ±0.23	6.61 ±1.75	8.08 ±0.47	10.03 ±0.80
WPZ	3.64 ±0.08	2.00 ±0.02	0.92 ±0.04	0.66 ±0.03	1.80 ±0.04	4.09 ±0.16	3.19 ±0.13	0.86 ±0.23	0.93 ±0.14	2.27 ±0.17
WPZ+GloVe	9.40 ±0.06	3.23 ±0.01	2.29 ±0.02	2.56 ±0.01	4.37 ±0.03	16.94 ±0.10	5.33 ±0.07	5.53 ±0.12	3.54 ±0.03	7.83 ±0.08
WPZ+BERT	32.49 ±2.01	3.89 ±0.24	10.68 ±1.40	6.67 ±0.46	13.43 ±1.03	**50.06** ±1.57	9.54 ±0.44	17.26 ±2.65	13.59 ±1.61	22.61 ±1.57
L-TapNet+CDT	**44.30** ±3.15	**12.04** ±0.65	**20.80** ±1.06	**15.17** ±1.25	**23.08** ±1.53	45.35 ±2.67	**11.65** ±2.34	**23.30** ±2.80	20.95 ±2.81	**25.31** ±2.65

Analysis

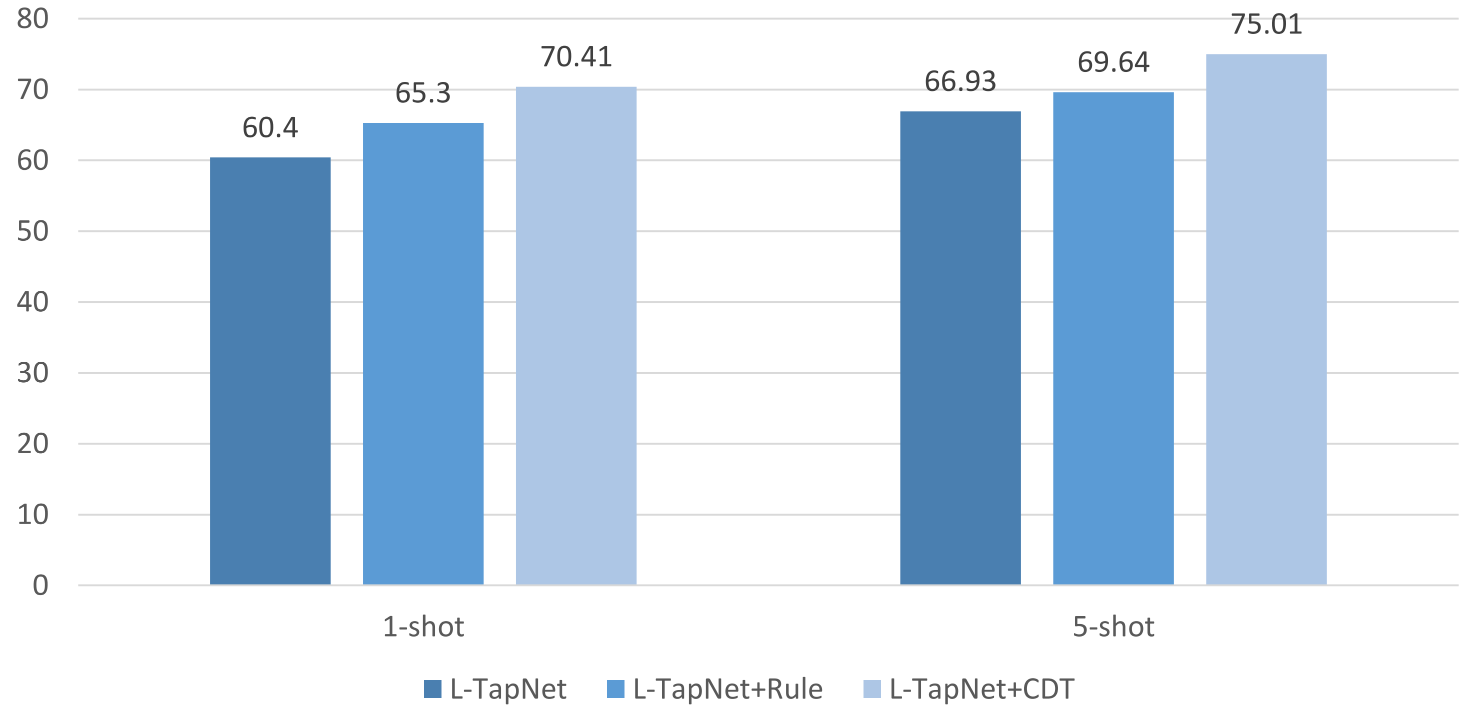
- How much does each module contribute?
  - We conduct ablation test and remove each component of our method respectively.

Model	1-shot	5-shots
Ours	70.41	75.01
- dependency transfer	-10.01	-8.08
- pair-wise embedding	-8.29	-7.74
- label semantic	-9.57	-4.87
- prototype reference	-1.73	-3.33

Ablation test results.

Analysis

- Does CDT just learn simple transition rules?



Comparison between transition rules and collapsed dependency transfer (CDT).

Case Analysis

- We analyzed the effectiveness of our method in different transition cases

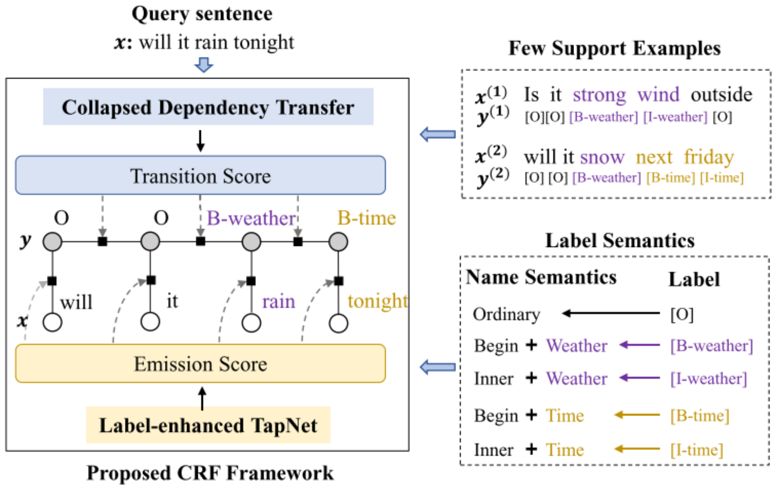
Bi-gram Type		Proportion	L-TapNet	+CDT
Border	O-O	28.5%	82.7%	**83.7%**
	O-B	24.5%	78.3%	**81.5%**
	B-O	8.2%	72.4%	**74.8%**
	I-O	5.8%	76.7%	**81.7%**
	I-B/B-B	7.8%	65.0%	**72.5%**
Inner	B-I	13.3%	78.5%	**83.6%**
	I-I	12.1%	77.8%	**82.7%**

- Border: CDT can implicitly help the model to decide slot boundaries
- Inner: CDT can learn the internal consistency of the slots

Thanks for listening!

Conclusion:

- We propose a CRF framework for Few-shot Slot Tagging.
- We propose the L-TapNet to leverage semantics of label names to enhance label representations
- We introduce the Collapsed Dependency Transfer to transfer label dependencies across domains with different label sets.
- We introduce Pair-wise Embedding to provide token representations.



Proposed CRF Framework