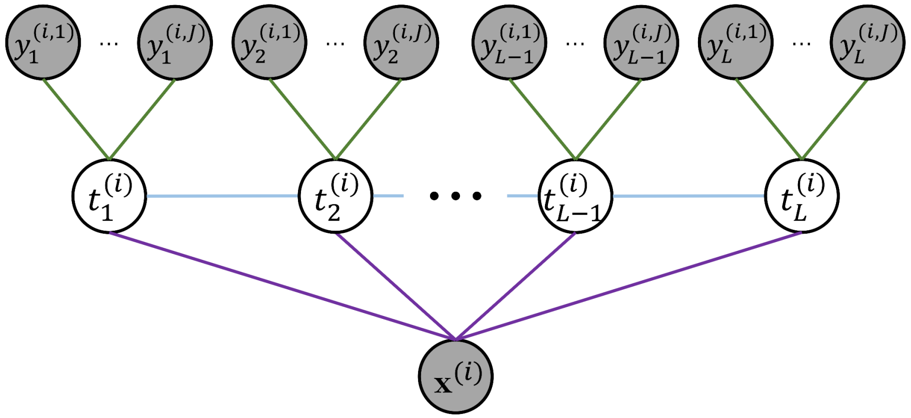


Neural-Hidden-CRF: A Robust Weakly-Supervised Sequence Labeler

Zhijun Chen[1], Hailong Sun[1], Wanhao Zhang[2], Chunyi Xu[1], Qianren Mao[3], Pengpeng Chen[4]

1 Beihang University, China | 2 Tsinghua University, China | 3 Zhongguancun Laboratory, China |
4 China's Aviation System Engineering Research Institute, China

Background

- How to learn from weakly-supervised sequence labels (to obtain a robust $f: X \rightarrow Y$)?
- An example of training data (on one datapoint):

Sentence X:	*"Jobs*	*returned*	*to*	*Apple*	*in*	*1997"*
Labels from weak supervision source #1 (crowdsourcing worker):	~~Others~~	Others	Others	Organization	Others	Others
Labels from weak supervision source #2 (domain rules and heuristics):	Person	Others	Others	~~Location~~	Others	Others
Labels from weak supervision source #3 (weak classifier):	Person	Others	Others	~~Location~~	Others	~~Miscellaneous~~
Truth Y (***unobserved***):	Person	Others	Others	Organization	Others	Others

Background

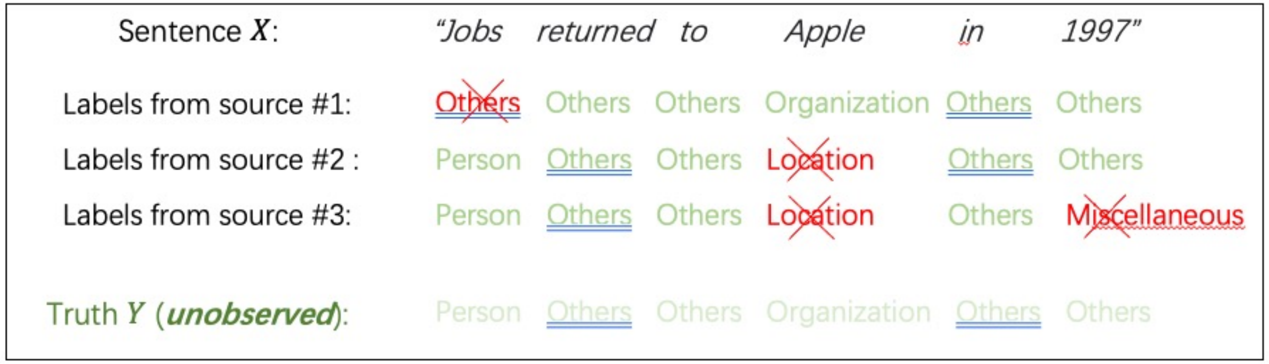
- Task: **Weak Supervision Sequence Labeling** (**WSSL**)
    - Sequence Labeling: Named Entity Recognition (NER), Part-of-speech (POS) Tagging…
    - Weak Supervision Learning
        - An alternative learning that is efficient, low-cost, and easy-to-promote on various domains
        - Labels can be contributed from various weak supervision sources

- WSSL is a widely studied problem [1]:
    - because of the importance of sequence labeling and weak supervision learning
    - because of the challenges associated with the need to take into account the abilities of diverse weak supervision sources and the back-and-forth dependence of the latent truth label sequence when solving the problem…

[1] Zhang et al. WRENCH: A Comprehensive Benchmark for Weak Supervision. NeurIPS 2021.

A very simple baseline

- Method: use the "Majority Voting" for truth inference → supervised sequence learning



Sentence X:	"*Jobs*	*returned*	*to*	*Apple*	*in*	*1997*"
Labels from source #1:	~~Others~~	Others	Others	Organization	Others	Others
Labels from source #2 :	Person	Others	Others	~~Location~~	Others	Others
Labels from source #3:	Person	Others	Others	~~Location~~	Others	~~Miscellaneous~~
Truth Y (***unobserved***):	Person	Others	Others	Organization	Others	Others

Results of MV (Majority Voting): Person Others Others ~~Location~~ Others Others

Just having the Majority Voting is not enough

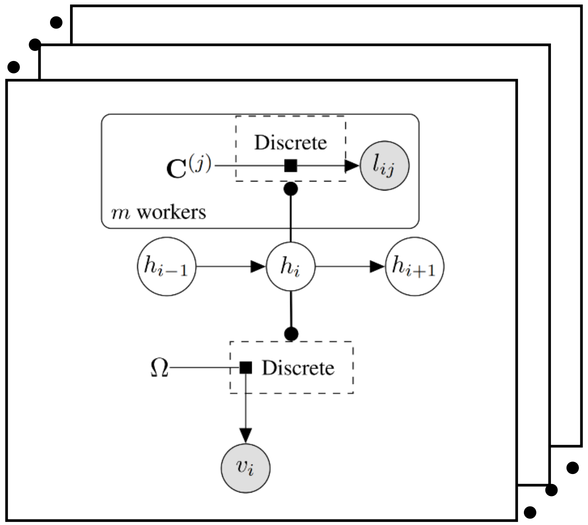
- Because this method is naïve. (For example, this method treats each weak supervision source equally without modelling their error rates or behavioral patterns.)
- More advanced methods have been proposed:

[2] Li et al. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. ACL 2022.
[3] Li et al. Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition. KDD 2022.
[4] Lan et al. Learning to contextually aggregate multi-source supervision for sequence labeling. ACL 2020.
[5] Zhang et al. Crowdsourcing Learning as Domain Adaptation: A Case Study on Named Entity Recognition. ACL 2021.
[6] Nguyen et al. Aggregating and predicting sequence labels from crowd annotations. ACL 2017.
[7] Lison et al. skweak: Weak Supervision Made Easy for NLP. ACL 2021.
[8] Lison et al. Named entity recognition without labelled data: A weak supervision approach. ACL 2020.
[9] Safranchik et al. Weakly supervised sequence tagging from noisy rules. AAAI 2020.
[10] Simpson et al. A Bayesian Approach for Sequence Tagging with Crowds. EMNLP 2019.
[11] Rodrigues et al. Deep learning from crowds. AAAI 2018.
[12] Sabetpour, et al. Optsla: an optimization based approach for sequential label aggregation. EMNLP 2020.
[13] Sabetpour et al. Truth discovery in sequence labels from crowds. ICDM 2021.
[14] Chen et al. Learning from Noisy Crowd Labels with Logics. ICDE 2023.
...

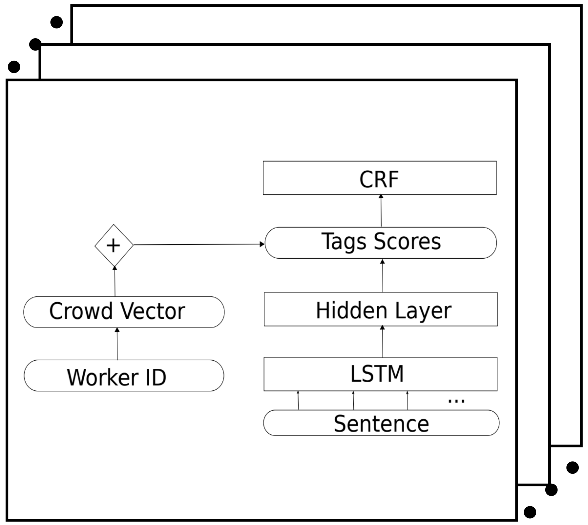
- Further, existing methods fall into two WSSL learning paradigms:
  - Two-stage: truth inference → supervised sequence learning
  - One-stage: perform the directly end-to-end learning to obtain a classifier

Existing representative methods

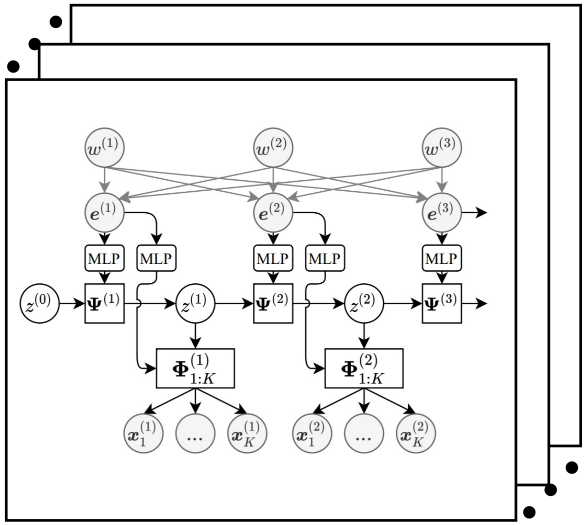
- Representative methods fall into three categories in intrinsic methodology:



1) HMM-based graphical models



2) "Source-specific perturbation" DL models



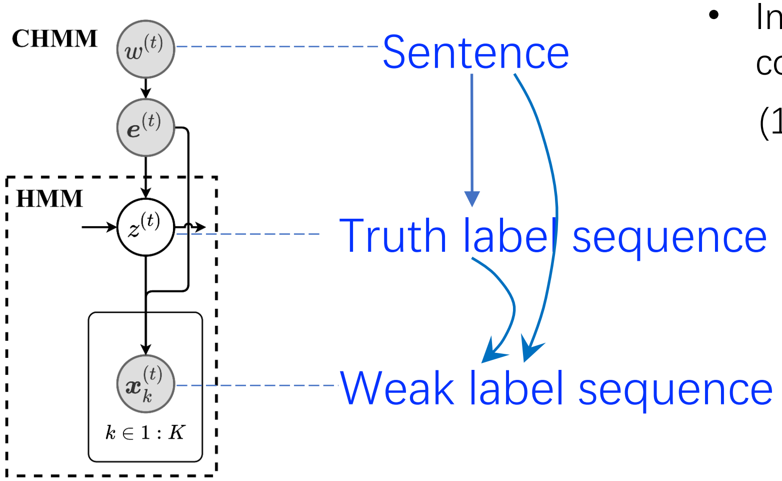
3) Neuralized HMM-based graphical models

- The Neuralized HMM-based graphical models have the advantages of both the first two approaches—i.e., the principled modeling of graphical models and the rich contextual knowledge that comes from DL models like BERT—and have achieved SOTA performances empirically [1]

[1] Zhang et al. WRENCH: A Comprehensive Benchmark for Weak Supervision. NeurIPS 2021.

The SOTA methods and drawbacks

- Neuralized HMM-based graphical models: CHMM [2], and its upgraded version Sparse-CHMM [3]. Because of their essential similarity, here we analyse the basic CHMM:



- In this directed probabilistic graphical model, three variables are involved and two dependencies are constructed:

  (1) The generation of the truth label sequence depends on the given sentence:

  - It uses the independence assumption to split the truth sequence into multiple regions and models the distribution of truth at each time step using a conditional model; i.e., it belongs to *per time-step modelling*: $p(z^{(t)}|z^{(t-1)}, Sentence; \Theta)$
  - It models "per-time-step-scaled patterns", i.e., the dependances concerning the current truth $z^{(t)}$ given the last truth $z^{(t-1)}$ and the sentence (obtaining local knowledge, using the local optimization perspective), rather than the dependances concerning the whole truth label sequence given the sentence (obtaining global knowledge, using the global optimization perspective)
  - *This per time-step modelling approach, like the well-known supervised learning method MEMM [15], applies the local optimization perspective and will inevitably lead to the well-known Label Bias Problem (where the sequence of labels predicted by the model has a certain bias); a wealth of analyses and proofs already exist [16,17,18,19]*

  (2) The generation of the weak label sequence depends on the truth label sequence and the sentence

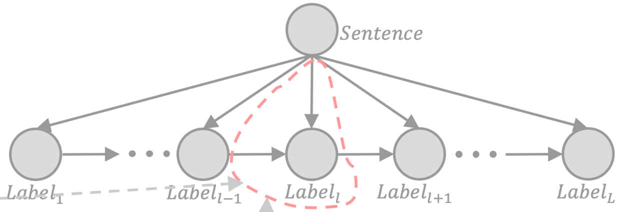
[2] Li et al. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. ACL 2022.
[3] Li et al. Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition. KDD 2022.
[15] McCallum et al. Maximum entropy Markov models for information extraction and segmentation. ICML 2000.
[16] John et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
[17] Hannun, Awni. The label bias problem. 2020.
[18] Charles Sutton, Andrew McCallum. An introduction to conditional random fields. Foundations and Trends® in Machine Learning, 2012.
[19] Simoes et al. Information Extraction tasks: a survey. Simpósio de Informática 2009.

Diving into the drawbacks

- Applies the local optimization perspective：
  - Performs the per time-step modeling:

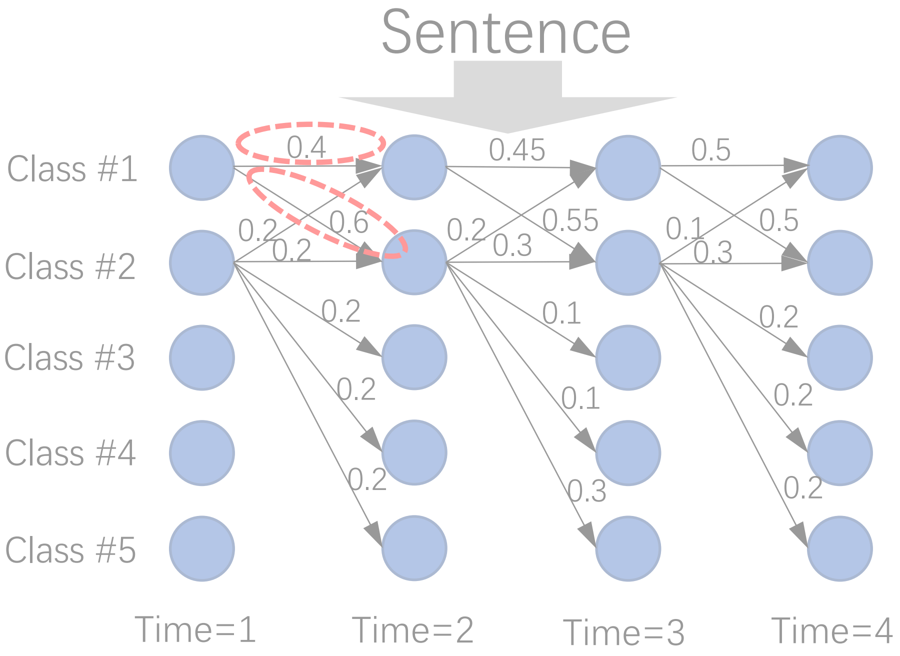
$$p(Label\_sequence|Sentence;\ \Theta) = \prod_{l} p(Label_l|Label_{l-1},\ Sentence;\Theta)$$

$$p(Label_l|Label_{l-1}, Sentence;\Theta) = \frac{\exp(f_\Theta(Label_l, Label_{l-1}, Sentence))}{\Sigma_{Truth_l}\exp(f_\Theta(Label_l, Label_{l-1}, Sentence))}$$



  - Model (i.e., Θ) repeatedly considers the local pattern for the scale of a time-step, obtains local knowledge

- Leads to the well-known Label Bias Problem
  - Background: During the prediction phase, the model needs to find the optimal label sequence, $Label\_sequence^* = \text{argmax}_{Label\_sequence} p(Label\_sequence|Sentence;\Theta)$; this per time-step modeling approach will result in a score map with restricted values (based on the strict conditional probability form)
  - Such restricted scores (e.g., 0.4, 0.6), as opposed to flexible scores (e.g., 0.2, 1.5), will be unfavourable, and will make the predicted label sequence has some bias (a bias toward states with fewer outgoing transitions) [17,18]



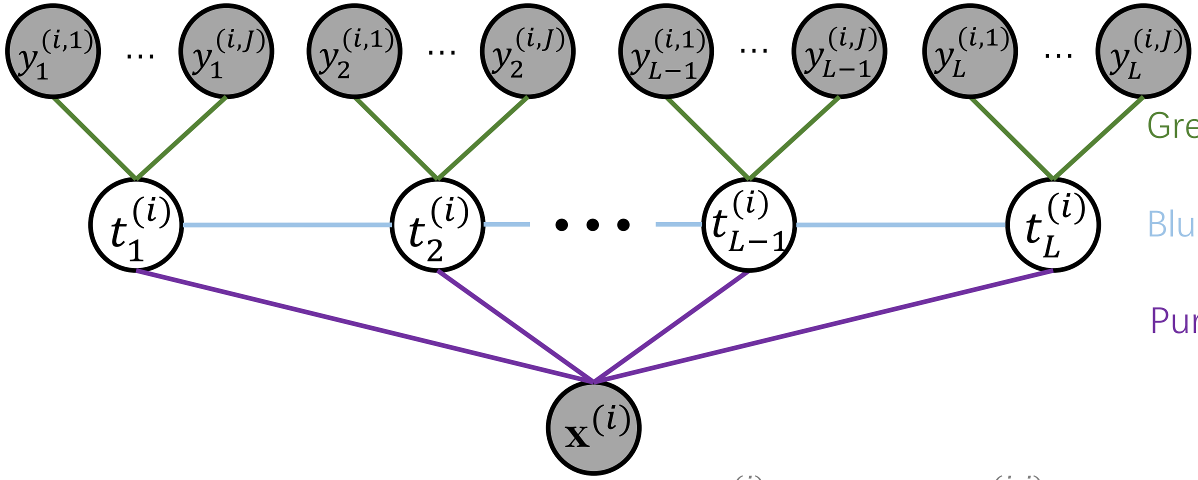
[17] Hannun, Awni. The label bias problem. 2020.
[20] Eric Xing. Lecture of Probabilistic Graphical Models. 2020.

Motivation

- Whether/how we can build a WSSL model:
  - capitalize on the graphical model with principled modeling of variable dependencies
  - capitalize on the advanced deep learning model that can bring rich contextual knowledge
  - without introducing the label bias problem caused by the local optimization perspective

Our Neural-Hidden-CRF

- Our model is simple: neuralized undirected graphical model **Neural-Hidden-CRF**:


$y_1^{(i,1)}$ ... $y_1^{(i,J)}$
$y_2^{(i,1)}$ ... $y_2^{(i,J)}$
$y_{L-1}^{(i,1)}$ ... $y_{L-1}^{(i,J)}$
$y_L^{(i,1)}$ ... $y_L^{(i,J)}$
$t_1^{(i)}$
$t_2^{(i)}$
• • •
$t_{L-1}^{(i)}$
$t_L^{(i)}$
$\mathbf{x}^{(i)}$

Green: J weak source transition matrices

Blue: the one CRF transition matrix

Purple: deep model, e.g., BERT

Figure: Neural-Hidden-CRF ($\mathbf{x}^{(i)}$: sentence; $t_l^{(i)}$: truth label; $y_l^{(i,j)}$: weak label)

- It reminds us of the similar classical models CRF and Neural-CRF (e.g., BERT-CRF) for supervised sequence learning:

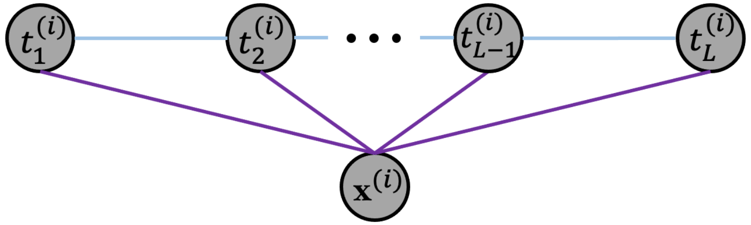


Figure: CRF/BERT-CRF ($\mathbf{x}^{(i)}$: sentence; $t_l^{(i)}$: truth label)

Recall classical models BERT-CRF/CRF

- BERT-CRF's Model:
  - $p(\mathbf{t}|\mathbf{x}) = \dfrac{\exp(\mathrm{score}_\Theta(\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{t}} \exp(\mathrm{score}_\Theta(\mathbf{t},\mathbf{x}))}$ (softmax on $\mathrm{score}_\Theta(\mathbf{t},\mathbf{x})$)

    (omit (i))

$$\mathrm{score}_\Theta(\mathbf{t},\mathbf{x}) = \sum_{l=1}^{L} [\mathrm{Emission}]_{l,t_l} + \sum_{l=1}^{L} [\mathrm{CrfTransition}]_{t_{l-1},t_l}$$

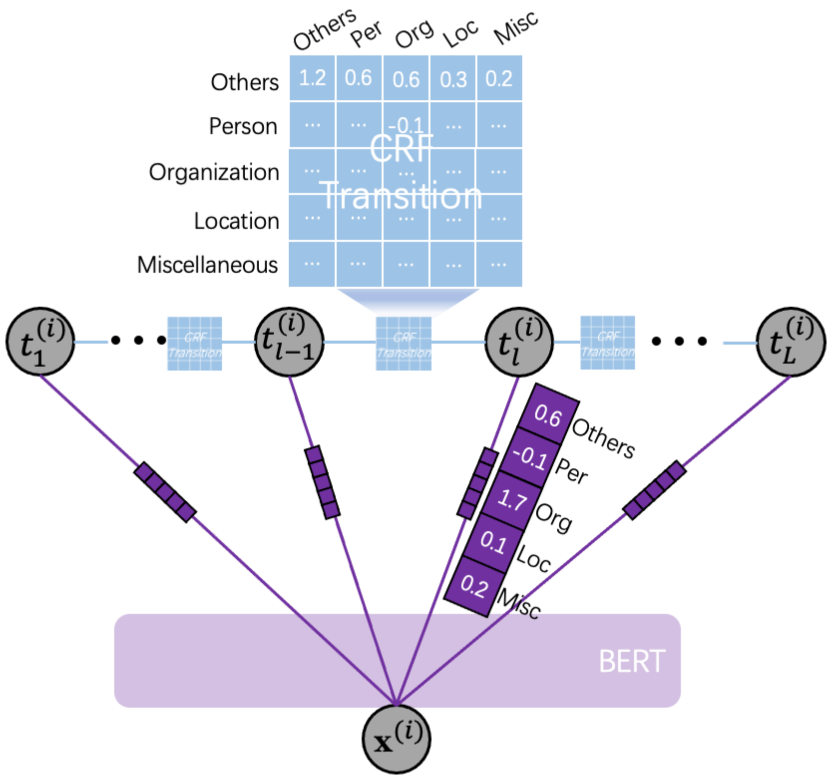
$$= \sum_{l=1}^{L} \left([f_{\mathrm{BERT}}\,(\mathbf{x};\Theta_{\mathrm{BERT}})]_{l,t_l} + [\mathrm{CrfTransition}]_{t_{l-1},t_l} \right)$$

Learned parameters $\Theta = \{\Theta_{\mathrm{BERT}}, [\mathrm{CrfTransition}]\}$

> *Example:*
> For t_{l-1} = "Others", t_l = "Organization", $\mathbf{x}$=*"Jobs returned to Apple in 1997"*, l =4:
> $[f_{\mathrm{BERT}}\,(\mathbf{x};\Theta_{\mathrm{BERT}})]_{l,t_l} + [\mathrm{CrfTransition}]_{t_{l-1},t_l} = 1.7 + 0.6 = 2.3$

- BERT-CRF's Objective and prediction
  - $\log p(\mathbf{t}^{(i)}|\mathbf{x}^{(i)})$ (efficiently compute with the Dynamic Programming [16])
  - Find the best sequence $\mathbf{t}^*$, $\mathbf{t}^* = \mathrm{argmax}_{\mathbf{t}}\, p(\mathbf{t}|\mathbf{x};\Theta)$

- CRF is similar: $[\mathrm{Emission}]_{l,t_l} = [\Theta_{\mathrm{CrfEmission}}]_{x_l,\,t_l}$



	Others	Per	Org	Loc	Misc
Others	1.2	0.6	0.6	0.3	0.2
Person	···	···	-0.1	···	···
Organization	···	···	···	···	···
Location	···	···	···	···	···
Miscellaneous	···	···	···	···	···

Figure: BERT-CRF ($\mathbf{x}^{(i)}$: sentence, $t_l^{(i)}$: truth label at l-th time step)

[16] John et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Our Neural-Hidden-CRF

- BERT-CRF:
  - $p(\mathbf{t}|\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{t}} \exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}$

    $$\text{score}_\Theta\ (\mathbf{t},\mathbf{x}) = \sum_{l=1}^{L}([\text{Emission}]_{l,t_l}+[\text{CrfTransition}]_{t_{l-1},t_l})$$

- Neural-Hidden-CRF:
  - $p(\mathbf{y},\mathbf{t}|\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}} \exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$

    $$\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x})) = \sum_{l=1}^{L}([\text{Emission}]_{l,t_l} + [\text{CrfTransition}]_{t_{l-1},t_l} + \uparrow\ )$$

    $$[\text{WeakSourceTransition\#1}]_{t_l,y^{(i,1)}} + \cdots + [\text{WeakSourceTransition\#}J]_{t_l,y^{(i,J)}}$$

Parameters $\Theta = \{\Theta_{\text{BERT}}, [\text{CrfTransition}], [\text{WeakSourceTransition\#1}], \ldots, [\text{WeakSourceTransition\#}J]\ \}$

- Objective: $\log p(\mathbf{y}|\mathbf{x}) = \log \sum_{\mathbf{t}} p(\mathbf{y},\mathbf{t}|\mathbf{x}) = \log \frac{\sum_{\mathbf{t}} \exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}} \exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$

  (efficiently compute with the Dynamic Programming like the CRFs)

- Prediction: use the classifier part (BERT-CRF) to find the best sequence $\mathbf{t}^* = \text{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{x};\Theta)$

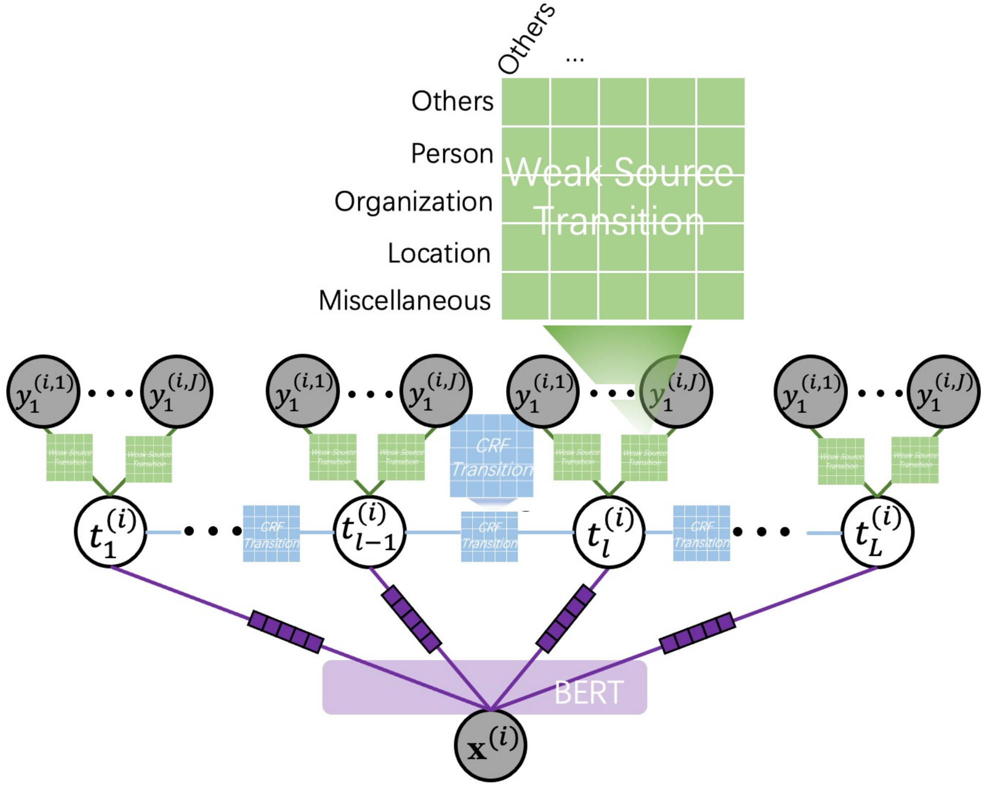
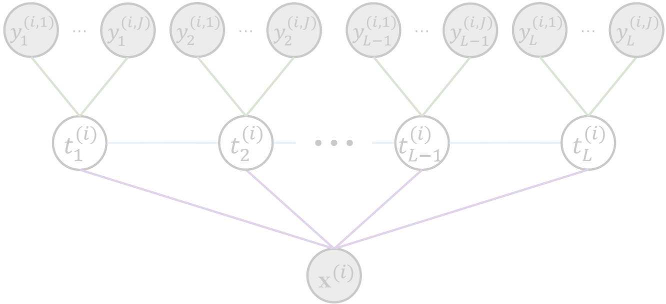


Figure: Neural-Hidden- CRF

($\mathbf{x}^{(i)}$: sentence, $t_l^{(i)}$: truth label at l-th time step)

Methodological advantages

- Benefits both from the principled modeling of graphical models, and from contextual knowledge of deep learning models
- Adopts the global normalization approach, avoiding the Label Bias Problem caused by the per time-step modeling approach
  - Global normalization (applying the global optimization perspective):
    - Modeling: $p(\mathbf{y}, \mathbf{t}|\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{y}, \mathbf{t}, \mathbf{x}))}{\sum_{\mathbf{y}} \sum_{\mathbf{t}} \exp(\text{score}_\Theta(\mathbf{y}, \mathbf{t}, \mathbf{x}))}$
    - Model (i.e., Θ) holistically considers global patterns for the entire truth sequence and weak label sequence given the sentence, obtains global knowledge



  - Avoids the Label Bias Problem (LBP):
    - This is because the adoption of the global normalization approach (which yields flexible path scores, e.g. 0.2, 1.5 instead of 0.4,0.6) instead of the per time-step modeling approach (which is the direct cause of LBP [17]), i.e.,

$$p(Truth_l|Truth_{l-1}, Sentence; \Theta)= \frac{\exp(f_\Theta(Truth_l, Truth_{l-1}, Sentence))}{\sum_{Truth_l} \exp(f_\Theta(Truth_l, Truth_{l-1}, Sentence))}$$

    - CRF [16] vs. MEMM [15] ≈ Neural-Hidden-CRF vs. CHMMs [2,3]

[2] Li et al. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. ACL 2022.
[3] Li et al. Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition. KDD 2022.
[15] McCallum et al. Maximum entropy Markov models for information extraction and segmentation. ICML 2000.
[16] John et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
[17] Hannun, Awni. The label bias problem. 2020.

BERT-CRF/CRF, Neural-Hidden-CRF are Undirected Graphical Models (UGMs)

- Theory of UGMs (the joint probability distribution is the normalized product value of *potential functions*):
  - Have 3 maximum cluster
  - $p(X) = \frac{1}{\text{Normalization } Z=\sum_X \text{SCORE}(X)} \text{SCORE}(X), \ \ \text{SCORE}(X) = \prod_C \phi_C(X_C)$

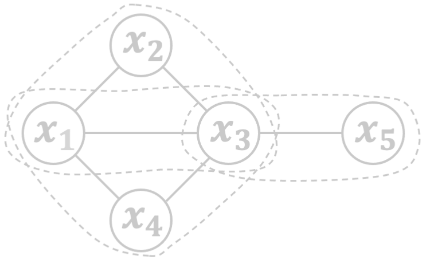


Figure: An example of UGMS

- BERT-CRF/CRF, Neural-Hidden-CRF are UGMs
  - BERT-CRF, CRF: For each time step, 1 maximum cluster
  - Neural-Hidden-CRF: For each time step, $1 + J$ maximum cluster

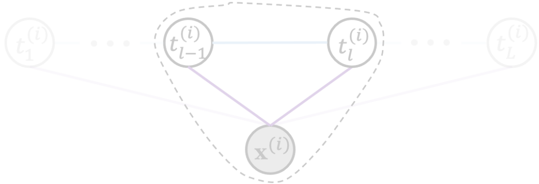


Figure: BERT-CRF/CRF

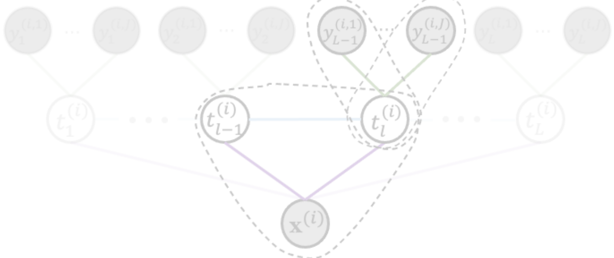


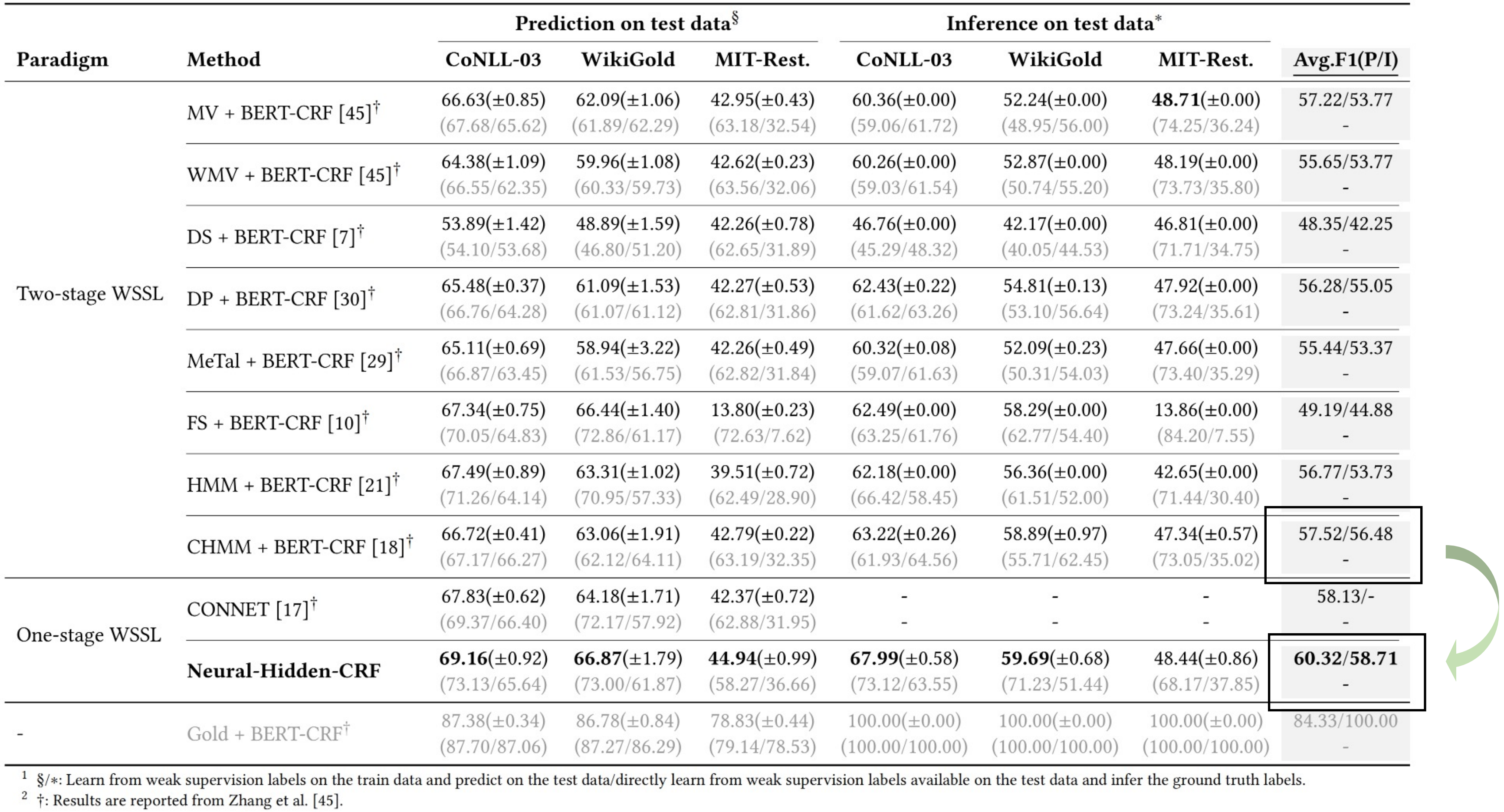
Figure: Neural-Hidden-CRF

Others

- Implementation Details: Parameter initialization
- Others: Computational Complexity
- Refer to the paper for more information

Results on three WS datasets [1]

- Prediction on test data / Inference on test data (exceeds CHMM by 2.80/2.23 F1)



		Prediction on test data§			Inference on test data*			
Paradigm	**Method**	**CoNLL-03**	**WikiGold**	**MIT-Rest.**	**CoNLL-03**	**WikiGold**	**MIT-Rest.**	**Avg.F1(P/I)**
Two-stage WSSL	MV + BERT-CRF [45]†	66.63(±0.85) (67.68/65.62)	62.09(±1.06) (61.89/62.29)	42.95(±0.43) (63.18/32.54)	60.36(±0.00) (59.06/61.72)	52.24(±0.00) (48.95/56.00)	**48.71**(±0.00) (74.25/36.24)	57.22/53.77 -
	WMV + BERT-CRF [45]†	64.38(±1.09) (66.55/62.35)	59.96(±1.08) (60.33/59.73)	42.62(±0.23) (63.56/32.06)	60.26(±0.00) (59.03/61.54)	52.87(±0.00) (50.74/55.20)	48.19(±0.00) (73.73/35.80)	55.65/53.77 -
	DS + BERT-CRF [7]†	53.89(±1.42) (54.10/53.68)	48.89(±1.59) (46.80/51.20)	42.26(±0.78) (62.65/31.89)	46.76(±0.00) (45.29/48.32)	42.17(±0.00) (40.05/44.53)	46.81(±0.00) (71.71/34.75)	48.35/42.25 -
	DP + BERT-CRF [30]†	65.48(±0.37) (66.76/64.28)	61.09(±1.53) (61.07/61.12)	42.27(±0.53) (62.81/31.86)	62.43(±0.22) (61.62/63.26)	54.81(±0.13) (53.10/56.64)	47.92(±0.00) (73.24/35.61)	56.28/55.05 -
	MeTal + BERT-CRF [29]†	65.11(±0.69) (66.87/63.45)	58.94(±3.22) (61.53/56.75)	42.26(±0.49) (62.82/31.84)	60.32(±0.08) (59.07/61.63)	52.09(±0.23) (50.31/54.03)	47.66(±0.00) (73.40/35.29)	55.44/53.37 -
	FS + BERT-CRF [10]†	67.34(±0.75) (70.05/64.83)	66.44(±1.40) (72.86/61.17)	13.80(±0.23) (72.63/7.62)	62.49(±0.00) (63.25/61.76)	58.29(±0.00) (62.77/54.40)	13.86(±0.00) (84.20/7.55)	49.19/44.88 -
	HMM + BERT-CRF [21]†	67.49(±0.89) (71.26/64.14)	63.31(±1.02) (70.95/57.33)	39.51(±0.72) (62.49/28.90)	62.18(±0.00) (66.42/58.45)	56.36(±0.00) (61.51/52.00)	42.65(±0.00) (71.44/30.40)	56.77/53.73 -
	CHMM + BERT-CRF [18]†	66.72(±0.41) (67.17/66.27)	63.06(±1.91) (62.12/64.11)	42.79(±0.22) (63.19/32.35)	63.22(±0.26) (61.93/64.56)	58.89(±0.97) (55.71/62.45)	47.34(±0.57) (73.05/35.02)	57.52/56.48 -
One-stage WSSL	CONNET [17]†	67.83(±0.62) (69.37/66.40)	64.18(±1.71) (72.17/57.92)	42.37(±0.72) (62.88/31.95)	- -	- -	- -	58.13/- -
	Neural-Hidden-CRF	**69.16**(±0.92) (73.13/65.64)	**66.87**(±1.79) (73.00/61.87)	**44.94**(±0.99) (58.27/36.66)	**67.99**(±0.58) (73.12/63.55)	**59.69**(±0.68) (71.23/51.44)	48.44(±0.86) (68.17/37.85)	**60.32/58.71** -
-	Gold + BERT-CRF†	87.38(±0.34) (87.70/87.06)	86.78(±0.84) (87.27/86.29)	78.83(±0.44) (79.14/78.53)	100.00(±0.00) (100.00/100.00)	100.00(±0.00) (100.00/100.00)	100.00(±0.00) (100.00/100.00)	84.33/100.00 -

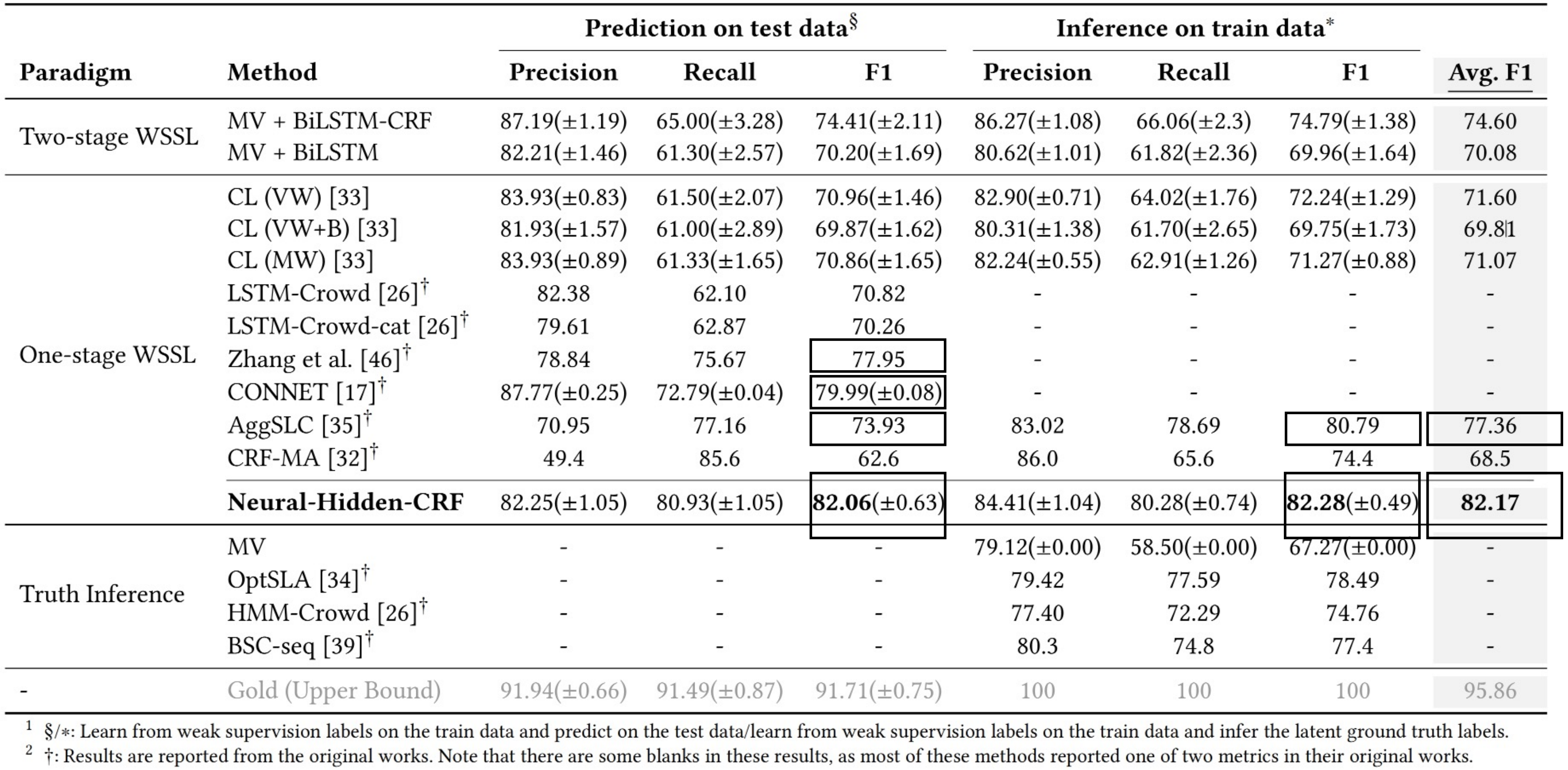
[1] §/*: Learn from weak supervision labels on the train data and predict on the test data/directly learn from weak supervision labels available on the test data and infer the ground truth labels.

[2] †: Results are reported from Zhang et al. [45].

[1] Zhang et al. WRENCH: A Comprehensive Benchmark for Weak Supervision. NeurIPS 2021.

Results on CoNLL-03 (MTurk) dataset [7]

- Prediction on test data / Inference on train data



		Prediction on test data§			Inference on train data*			
Paradigm	**Method**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Avg. F1**
Two-stage WSSL	MV + BiLSTM-CRF	87.19(±1.19)	65.00(±3.28)	74.41(±2.11)	86.27(±1.08)	66.06(±2.3)	74.79(±1.38)	74.60
	MV + BiLSTM	82.21(±1.46)	61.30(±2.57)	70.20(±1.69)	80.62(±1.01)	61.82(±2.36)	69.96(±1.64)	70.08
One-stage WSSL	CL (VW) [33]	83.93(±0.83)	61.50(±2.07)	70.96(±1.46)	82.90(±0.71)	64.02(±1.76)	72.24(±1.29)	71.60
	CL (VW+B) [33]	81.93(±1.57)	61.00(±2.89)	69.87(±1.62)	80.31(±1.38)	61.70(±2.65)	69.75(±1.73)	69.81
	CL (MW) [33]	83.93(±0.89)	61.33(±1.65)	70.86(±1.65)	82.24(±0.55)	62.91(±1.26)	71.27(±0.88)	71.07
	LSTM-Crowd [26]†	82.38	62.10	70.82	-	-	-	-
	LSTM-Crowd-cat [26]†	79.61	62.87	70.26	-	-	-	-
	Zhang et al. [46]†	78.84	75.67	77.95	-	-	-	-
	CONNET [17]†	87.77(±0.25)	72.79(±0.04)	79.99(±0.08)	-	-	-	-
	AggSLC [35]†	70.95	77.16	73.93	83.02	78.69	80.79	77.36
	CRF-MA [32]†	49.4	85.6	62.6	86.0	65.6	74.4	68.5
	Neural-Hidden-CRF	82.25(±1.05)	80.93(±1.05)	**82.06**(±0.63)	84.41(±1.04)	80.28(±0.74)	**82.28**(±0.49)	**82.17**
Truth Inference	MV	-	-	-	79.12(±0.00)	58.50(±0.00)	67.27(±0.00)	-
	OptSLA [34]†	-	-	-	79.42	77.59	78.49	-
	HMM-Crowd [26]†	-	-	-	77.40	72.29	74.76	-
	BSC-seq [39]†	-	-	-	80.3	74.8	77.4	-
-	Gold (Upper Bound)	91.94(±0.66)	91.49(±0.87)	91.71(±0.75)	100	100	100	95.86

[1] §/*: Learn from weak supervision labels on the train data and predict on the test data/learn from weak supervision labels on the train data and infer the latent ground truth labels.

[2] †: Results are reported from the original works. Note that there are some blanks in these results, as most of these methods reported one of two metrics in their original works.

[7] Rodrigues et al. Deep Learning from Crowds. AAAI 2018.

Weak Source Parameter Estimation

- Accurate weak source parameter estimation, and good interpretability
  - “Real” : true probabilistic confusion matrix; “Estimated by our (1)/(2)” : after normalization calculations on ours

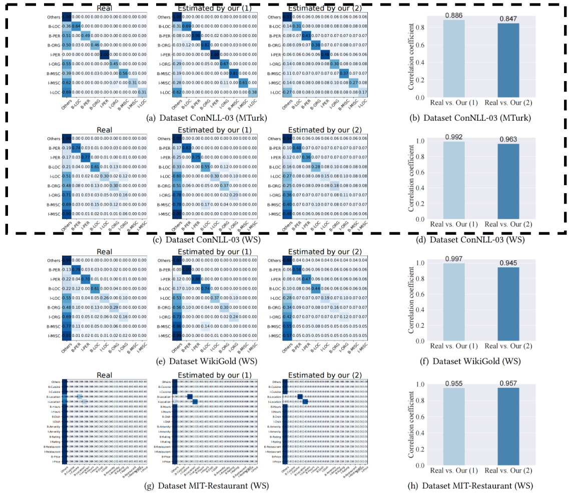
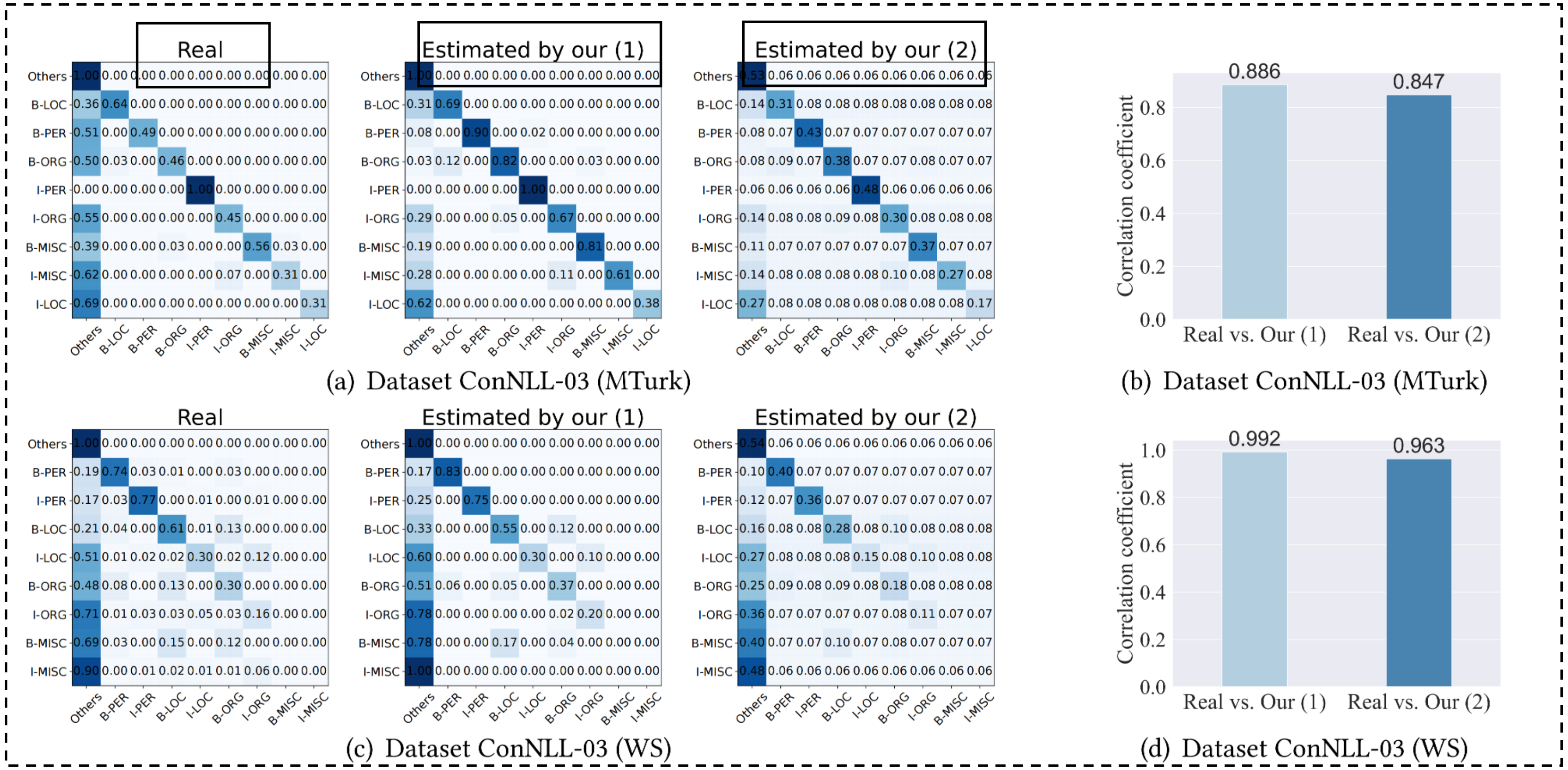


Figure: Results on four datasets.



(a) Dataset ConNLL-03 (MTurk)

(b) Dataset ConNLL-03 (MTurk)

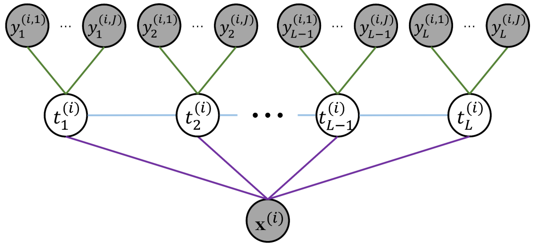
(c) Dataset ConNLL-03 (WS)

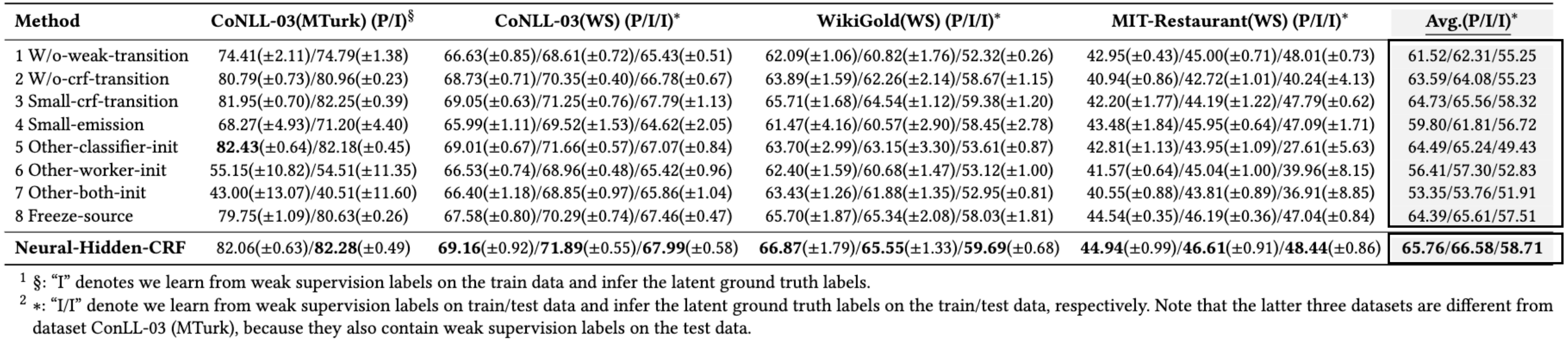
(d) Dataset ConNLL-03 (WS)

Figure: Results the the CoNLL-03 (MTurk) dataset and CoNLL-03 (WS) dataset.

Ablation Study

- Ablate the parts (weak source transition matrices (#1), CRF transition matrix (#2, #3), emission values (#4))
- Inappropriate parameter initialization (#5, #6, #7)
- Freeze model parameters (#8)





Method	CoNLL-03(MTurk) (P/I)§	CoNLL-03(WS) (P/I/I)*	WikiGold(WS) (P/I/I)*	MIT-Restaurant(WS) (P/I/I)*	Avg.(P/I/I)*
1 W/o-weak-transition	74.41(±2.11)/74.79(±1.38)	66.63(±0.85)/68.61(±0.72)/65.43(±0.51)	62.09(±1.06)/60.82(±1.76)/52.32(±0.26)	42.95(±0.43)/45.00(±0.71)/48.01(±0.73)	61.52/62.31/55.25
2 W/o-crf-transition	80.79(±0.73)/80.96(±0.23)	68.73(±0.71)/70.35(±0.40)/66.78(±0.67)	63.89(±1.59)/62.26(±2.14)/58.67(±1.15)	40.94(±0.86)/42.72(±1.01)/40.24(±4.13)	63.59/64.08/55.23
3 Small-crf-transition	81.95(±0.70)/82.25(±0.39)	69.05(±0.63)/71.25(±0.76)/67.79(±1.13)	65.71(±1.68)/64.54(±1.12)/59.38(±1.20)	42.20(±1.77)/44.19(±1.22)/47.79(±0.62)	64.73/65.56/58.32
4 Small-emission	68.27(±4.93)/71.20(±4.40)	65.99(±1.11)/69.52(±1.53)/64.62(±2.05)	61.47(±4.16)/60.57(±2.90)/58.45(±2.78)	43.48(±1.84)/45.95(±0.64)/47.09(±1.71)	59.80/61.81/56.72
5 Other-classifier-init	**82.43**(±0.64)/82.18(±0.45)	69.01(±0.67)/71.66(±0.57)/67.07(±0.84)	63.70(±2.99)/63.15(±3.30)/53.61(±0.87)	42.81(±1.13)/43.95(±1.09)/27.61(±5.63)	64.49/65.24/49.43
6 Other-worker-init	55.15(±10.82)/54.51(±11.35)	66.53(±0.74)/68.96(±0.48)/65.42(±0.96)	62.40(±1.59)/60.68(±1.47)/53.12(±1.00)	41.57(±0.64)/45.04(±1.00)/39.96(±8.15)	56.41/57.30/52.83
7 Other-both-init	43.00(±13.07)/40.51(±11.60)	66.40(±1.18)/68.85(±0.97)/65.86(±1.04)	63.43(±1.26)/61.88(±1.35)/52.95(±0.81)	40.55(±0.88)/43.81(±0.89)/36.91(±8.85)	53.35/53.76/51.91
8 Freeze-source	79.75(±1.09)/80.63(±0.26)	67.58(±0.80)/70.29(±0.74)/67.46(±0.47)	65.70(±1.87)/65.34(±2.08)/58.03(±1.81)	44.54(±0.35)/46.19(±0.36)/47.04(±0.84)	64.39/65.61/57.51
Neural-Hidden-CRF	82.06(±0.63)/**82.28**(±0.49)	**69.16**(±0.92)/**71.89**(±0.55)/**67.99**(±0.58)	**66.87**(±1.79)/**65.55**(±1.33)/**59.69**(±0.68)	**44.94**(±0.99)/**46.61**(±0.91)/**48.44**(±0.86)	**65.76/66.58/58.71**

[1] §: "I" denotes we learn from weak supervision labels on the train data and infer the latent ground truth labels.

[2] ∗: "I/I" denote we learn from weak supervision labels on train/test data and infer the latent ground truth labels on the train/test data, respectively. Note that the latter three datasets are different from dataset ConLL-03 (MTurk), because they also contain weak supervision labels on the test data.

Some other experiments

- Equipped with different backbones:
  - For the prediction task on datasets CoNLL-03 (WS) and WikiGold (WS), our Neural-Hidden-CRF: BiLSTM-based/BERT-based: 67.63(±1.08)/69.16(±0.92), 65.21(±1.45)/66.87(±1.79).
- More ablation studies:

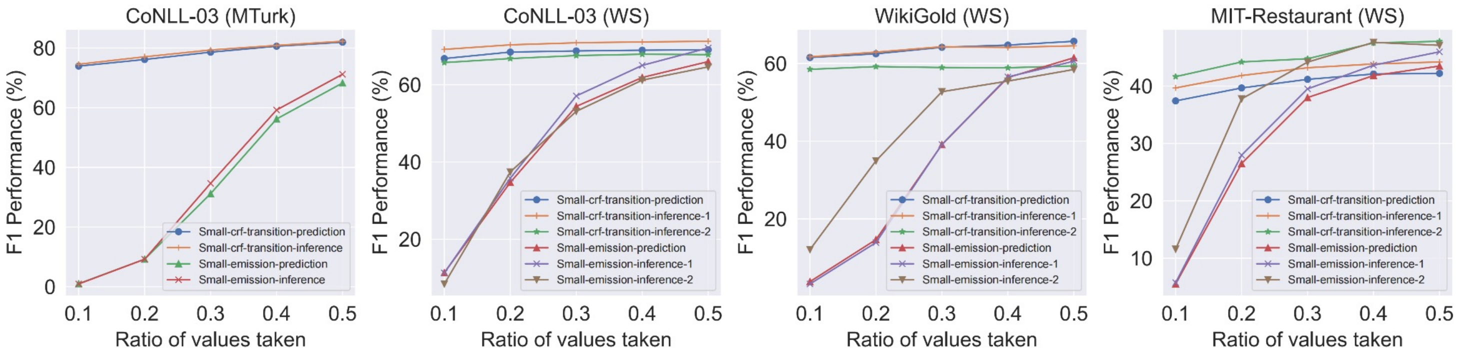
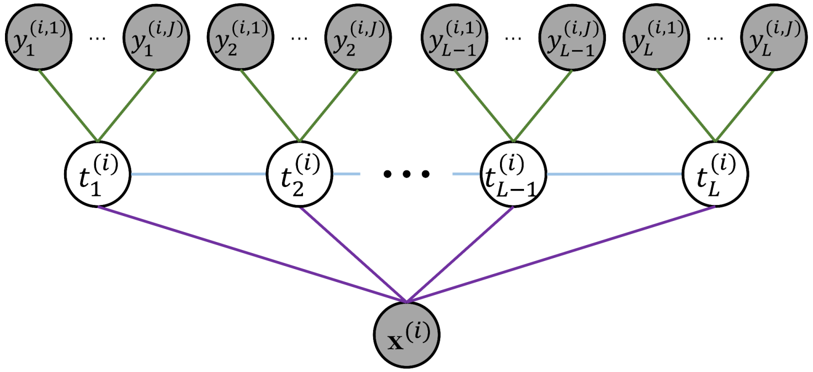


Figure A.1: Performance of more variants for supplementary ablation study. Results are averaged over 20 runs.

Conclusion

- For Learning from weakly-supervised sequence labels? Try Neural-Hidden-CRF
    - Neural-Hidden-CRF, the first neuralized undirected graphical model for the WSSL problem, benefits both from the principled modeling of graphical models and from contextual knowledge of deep learning models, while avoiding the label bias problem through the global optimization perspective
    - Code, and more information (slides): https://github.com/junchenzhi/Neural-Hidden-CRF


$y_1^{(i,1)}$
…
$y_1^{(i,J)}$
$y_2^{(i,1)}$
…
$y_2^{(i,J)}$
$y_{L-1}^{(i,1)}$
…
$y_{L-1}^{(i,J)}$
$y_L^{(i,1)}$
…
$y_L^{(i,J)}$
$t_1^{(i)}$
$t_2^{(i)}$
…
$t_{L-1}^{(i)}$
$t_L^{(i)}$
$\mathbf{x}^{(i)}$