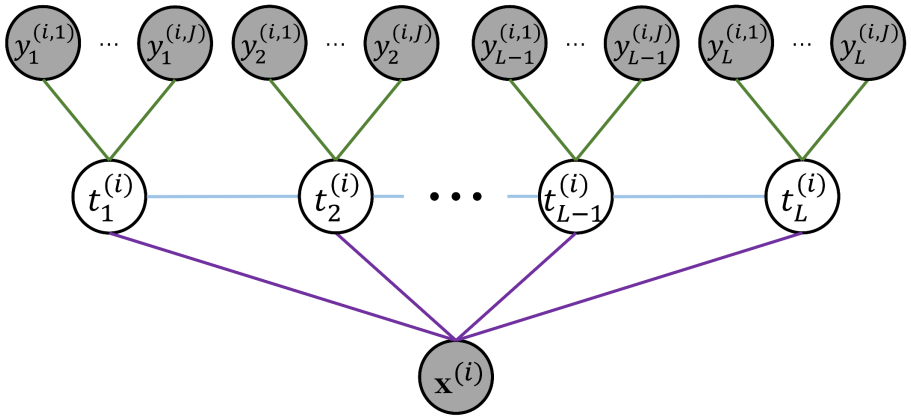


Neural-Hidden-CRF: 一个鲁棒的弱监督序列标注器

Neural-Hidden-CRF: A Robust Weakly-Supervised Sequence Labeler (KDD-2023)

报告人：陈志珺（导师：孙海龙）

BEIHANG UNIVERSITY ChineseCSCW

任务

- 如何利用 "弱监督序列标注数据" 来学习（以获得一个鲁棒的标注器，即分类器）？
- 训练数据的一个例子 (一个样本点):

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

任务

- 任务:**弱监督序列标注——Weak Supervision Sequence Labeling (WSSL)**
  - **序列标注** （Sequence Labeling）: 命名实体识别（NER）， 词性标注（POS tagging）…
  - **弱监督学习** （Weak Supervision Learning）
    - 流行的学习范式——快捷、低成本、在不同的领域中具有良好的可拓展性…
    - 标签可以来自不同的弱监督源：众包工人，人工定义函数…

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

面临挑战

- WSSL 是一个被广泛研究的问题 [1], 因为：
  - WSSL问题本身的重要性
  - 背后的挑战性
    - 多源异构的标注源（弱监督源的能力、行为模式不同）；如何刻画？
    - 真值标签序列具有标签前后依赖关系；如何刻画？

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

一个非常简单的方法

- 首先用 "大多数投票" （Majority Voting）方法来进行真值推理 → 再利用推理结果来进行有监督学习

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 Majority Voting (MV):	Person	Others	Others	Location	Others	Others

只拥有大多数投票方法还不够

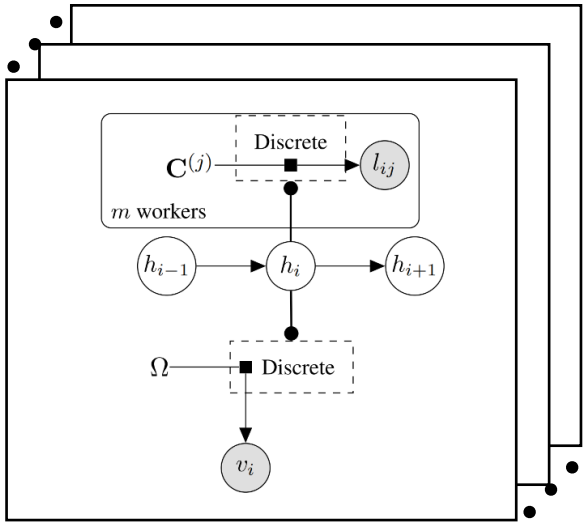
- 例如，这种方法平等地对待每个弱监督源，而不考虑它们所特有的错误率或行为模式
- 人们提出了更多更先进的方法：

[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.
...

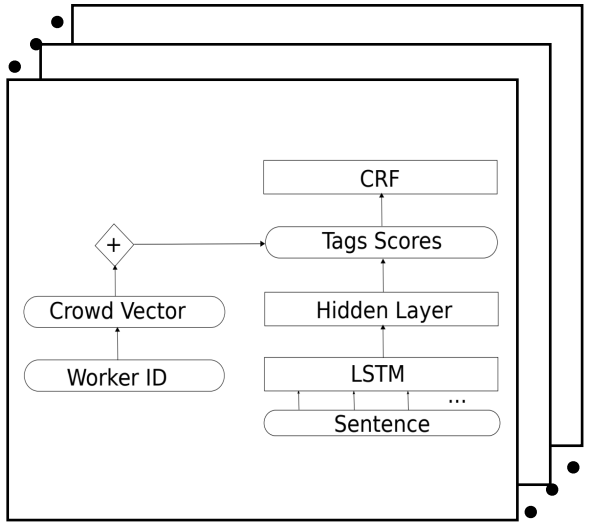
  - 它们都可划分为两类学习范式
    - 二阶段学习范式：先进行“真值推理” → 监督学习
    - 一阶段学习范式：直接利用弱标注，进行端对端学习以获得分类器

更先进的主流方法的划分

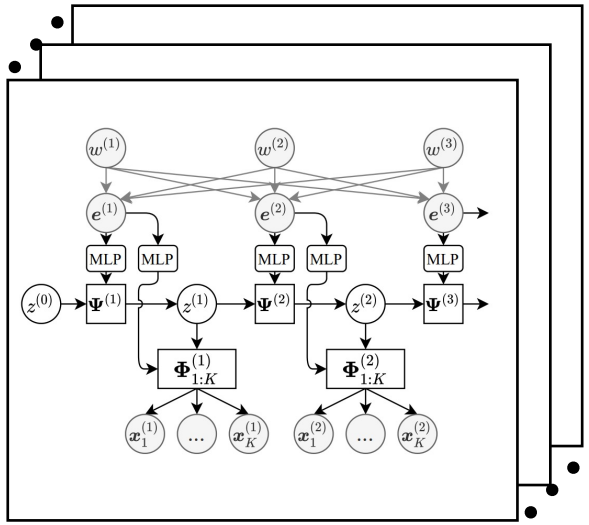
- 主流方法分为三类：



1) HMM-based graphical models



2) "Source-specific perturbation" DL models



3) Neuralized HMM-based graphical models

- 最近所提出的第三类方法（Neuralized HMM-based graphical models）：
  - 具有前两种方法的优势——即概率图模型的原则性建模，和来自BERT等深度学习模型的丰富的上下文知识
  - 并在实验中上实现了SOTA性能[1]

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

SOTA方法及其缺陷

- Neuralized HMM-based graphical models: CHMM [2] (ACL-2022), 和它的升级版本的方法 Sparse-CHMM [3] (KDD-2022)。因本质上的相似性，我们分析基础方法CHMM：
  - 在此有向概率图模型中，含有了3种变量，构建了2种依赖关系：
    - （1）**<u>真值标签序列的生成</u>**，依赖于给定的句子：
      - 利用独立性假设把真值序列拆分多个区域，对每个时间步上的真值标签的分布进行建模，即属于per time-step modeling：$p(z^{(t)}|z^{(t-1)}, Sentence; \Theta)$；
      - 它建模的"以一个时间步为规模的patterns"，即当前时刻的真值$z^{(t)}$所依赖于上一个时刻的真值$z^{(t-1)}$和$Sentence$的规律（即"局部知识"；局部优化视觉），而不是整个真值序列z依赖于$Sentence$的规律（即"全局知识"；全局优化视觉）

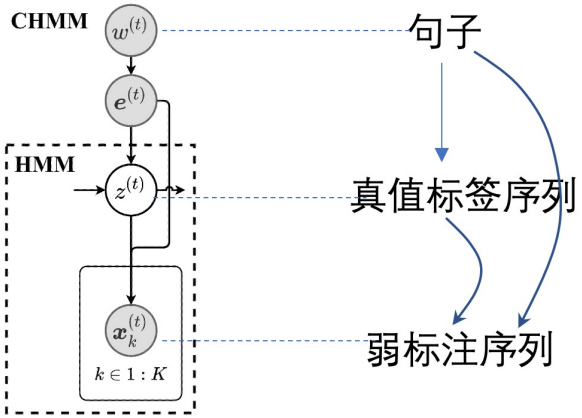


图: CHMM/Sparse-CHMM (在一个时间步上的示意图)

这种per time-step modeling，和著名的序列监督学习方法MEMM[15]一样，应用了一种局部优化视角，将不可避免地导致著名的Label Bias Problem（模型所预测的标签序列具有某种固定的偏差性）；已存在大量的分析和证明[16,17,18,19]

- （2）**<u>弱标注的生成</u>**，依赖于给定的句子和真值标签序列

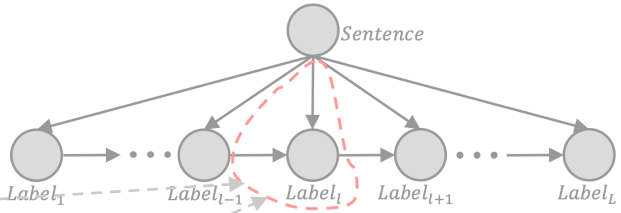
[2] Li et al. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. ACL 2022.
[3] Li et al. Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition. KDD 2022.
[15] McCallum et al. Maximum entropy Markov models for information extraction 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.

具体看SOTA方法的缺陷

- 应用局部优化视角：
  - 执行per time-step modeling:

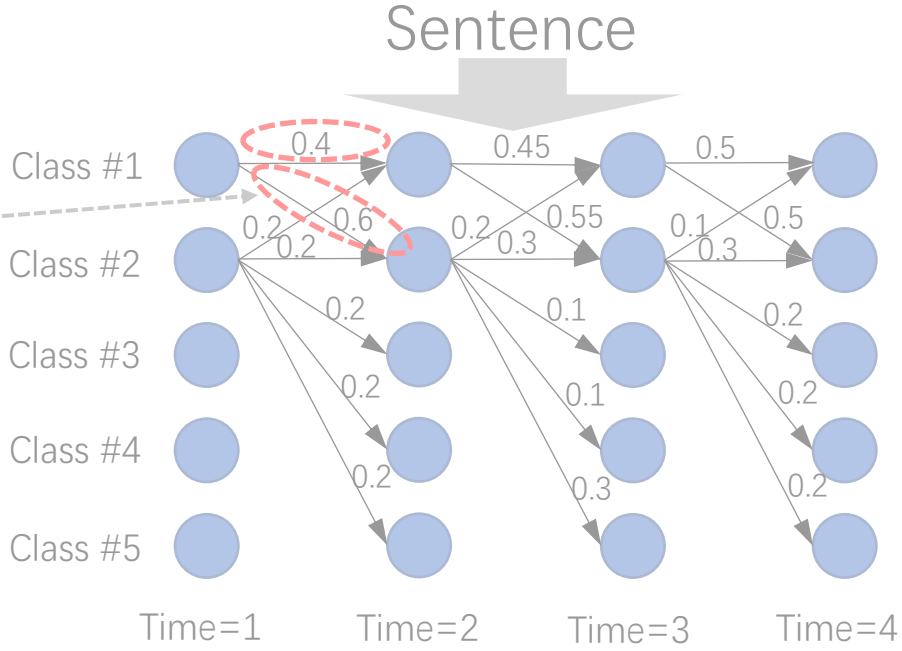
$$p(Label\_sequence|Sentence;\ \Theta) = \prod_{l} p(Label_l|Label_{l-1},\ Sentence;\Theta)$$

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



  - 模型 (即, Θ) 重复性考虑local patterns for the scale of a time-step, 获得local knowledge;

- 直接导致著名的Label Bias Problem (LBP)：
  - 背景：模型在预测阶段，需要寻找最优标签序列，$Label\_sequence^* = \text{argmax}_{Label\_sequence} p(Label\_sequence|Sentence;\Theta)$；这类per time-step modeling方法将形成（基于条件概率形式的）“具有限制性取值”的分数图
  - 这种“具有限制性取值”的分数 (比如0.4,0.6)，相对于“具有灵活性取值”的分数（比如0.2, 1.5)，将不利于我们，将会导致所求解出的标签序列具有偏差（“友谊链条”的例子[20]；基于“Information Erasure”的数学证明[17])



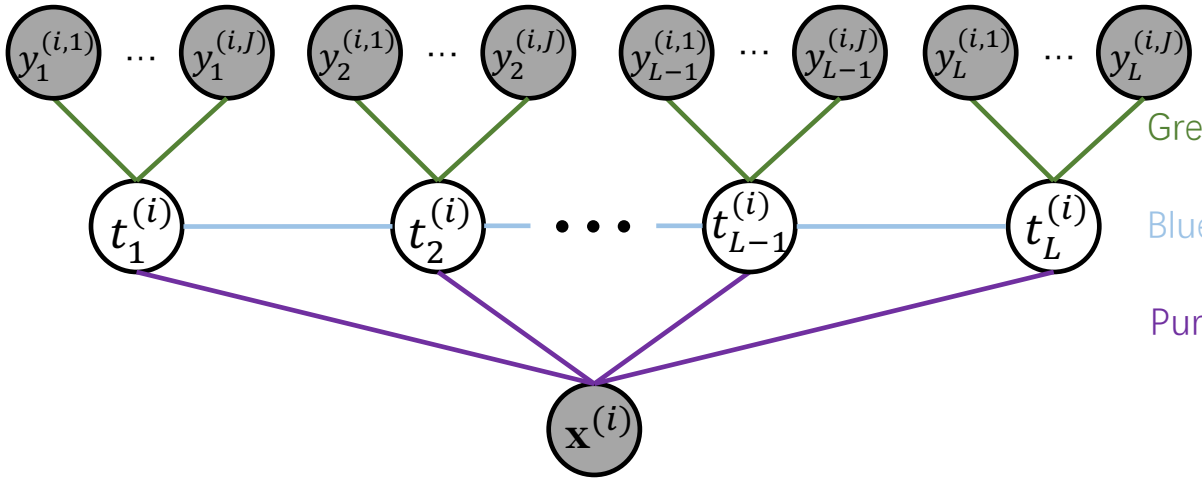
[17] Hannun, Awni. The label bias problem. 2020.
[20] Eric Xing. Lecture of Probabilistic Graphical Models. 2020.

研究动机

- 我们是否可以/如何构建一个WSSL模型：
  - 既利用——概率图模型所带来的对变量的有原则性建模；
  - 又利用——先进的深度学习模型（如BERT）所带来的丰富的上下文语义知识；
  - 同时不会——引入由局部优化视角所引起的Label Bias Problem？

Neural-Hidden-CRF

- 模型Neural-Hidden-CRF是一个神经化无向图模型；模型在图形表示上简单、对称：



Green: J weak source transition matrices

Blue: the one CRF transition matrix

Purple: deep model, e.g., BERT

图: Neural-Hidden-CRF ($\mathbf{x}^{(i)}$:句子, $t_l^{(i)}$: 在 l-th 时间步上的真值, $y_l^{(i,j)}$: 在 l-th时间步上的弱标注，来自弱标注源j)

- 类似于经典的监督学习算法CRF[16]和Neural-CRF(例如BERT-CRF) :

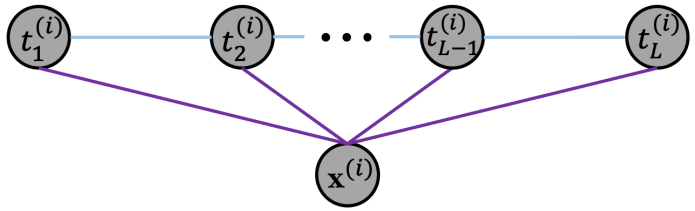


图: CRF/BERT-CRF ($\mathbf{x}^{(i)}$:句子, $t_l^{(i)}$: 在 l-th 时间步上的真值)

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

回忆经典模型BERT-CRF/CRF

- BERT-CRF的模型构建:
  - $p(\text{真值标签序列}\mathbf{t}|\text{句子}\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}$ (softmax on $\text{score}_\Theta(\mathbf{t},\mathbf{x})$)

(注：忽略 (i))

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

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

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

例子（计算$score_\Theta(t,x)$在一个时间步上的得分）：
对于 t_{l-1} = “Others”, t_l = “Organization”, $\mathbf{x}$=“*Jobs returned to Apple in 1997*”, l =4:
$[f_{\text{BERT}}(\mathbf{x};\Theta_{\text{BERT}})]_{l,t_l} + [\text{CrfTransition}]_{t_{l-1},t_l} = 1.7 + 0.6 = 2.3$

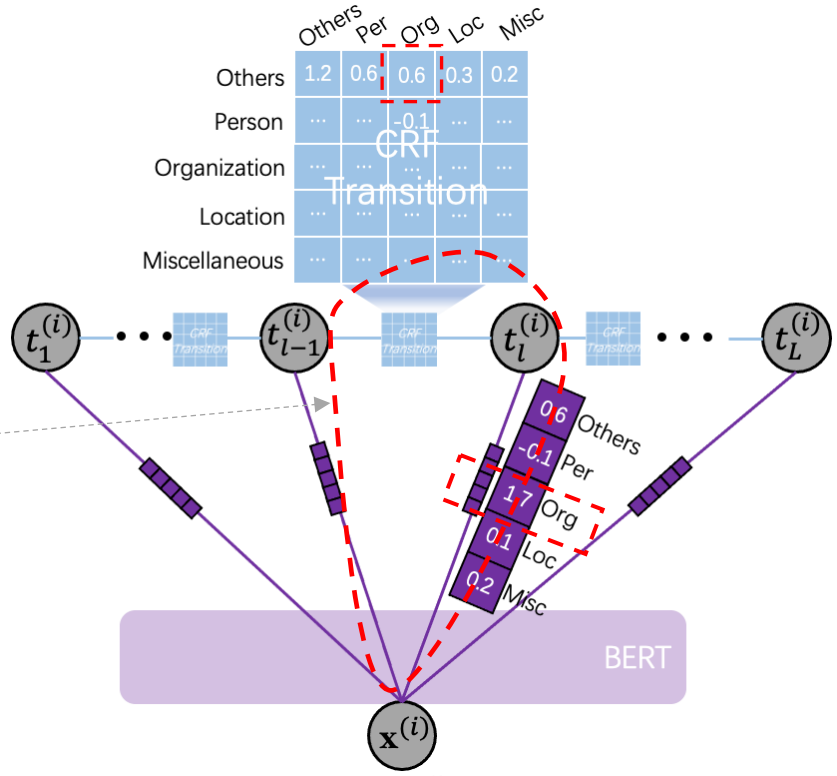


图: BERT-CRF ($\mathbf{x}^{(i)}$:句子, $t_l^{(i)}$:在 l-th时间步上的真值标签)

- BERT-CRF的优化目标/预测：
  - 目标：$\log p(\mathbf{t}^{(i)}|\mathbf{x}^{(i)})$ (用Dynamic Programming做快速计算[16])
  - 预测：寻找最优序列 $\mathbf{t}^* = \text{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{x};\Theta)$

- CRF是类似的(在公式中，把$[f_{\text{BERT}}(\mathbf{x};\Theta_{\text{BERT}})]_{l,t_l}$ 替换为$[\Theta_{\text{CrfEmission}}]_{x_l,t_l}$)

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

所提出的Neural-Hidden-CRF

- BERT-CRF:
  - 模型：$p(\text{真值标签序列}\mathbf{t}|\text{句子}\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{t},\mathbf{x}))}$

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

- Neural-Hidden-CRF:
  - 模型：$p(\text{弱标注序列}\mathbf{y}, \text{真值序列}\mathbf{t}|\text{句子}\mathbf{x}) = \frac{\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$

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

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

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

- 优化: $\log p(\mathbf{y}|\mathbf{x}) = \log\sum_{\mathbf{t}} p(\mathbf{y},\mathbf{t}|\mathbf{x}) = \log\frac{\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$
  (类似BERT-CRF，可用Dynamic Programming做快速计算)
- 预测: 用分类器(如BERT-CRF) 来寻找最优序列$\mathbf{t}^* = \text{argmax}_{\mathbf{t}}p(\mathbf{t}|\mathbf{x};\Theta)$

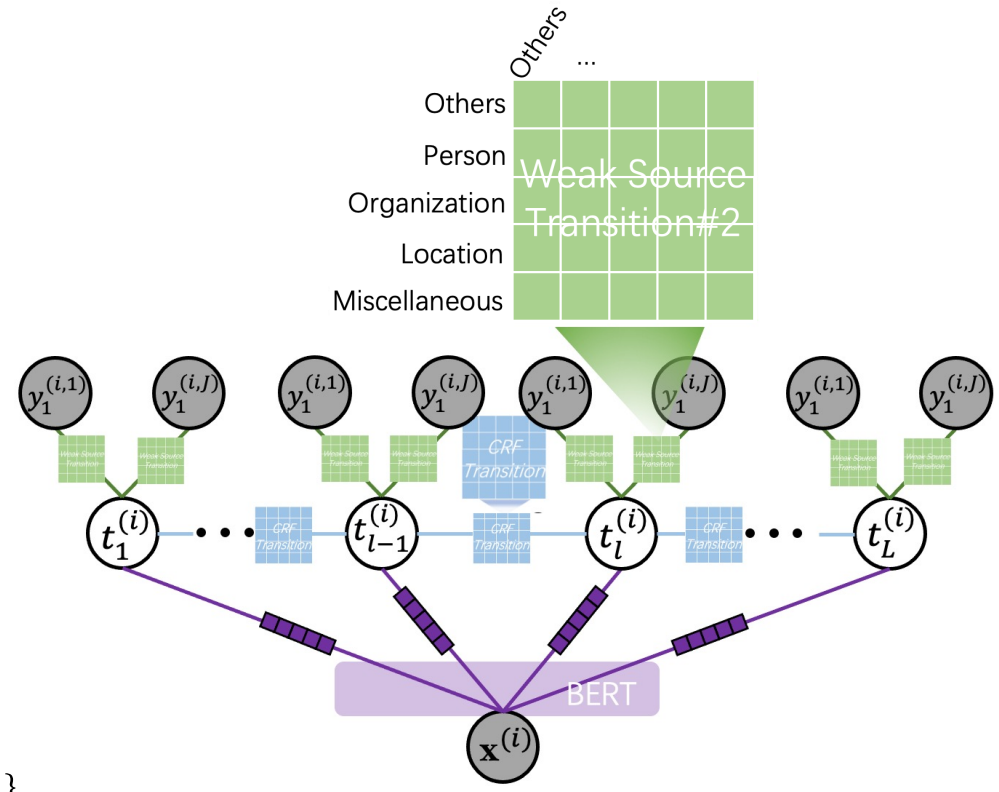


图: Neural-Hidden- CRF
($\mathbf{x}^{(i)}$:句子，$t_l^{(i)}$: 在 l-th时间步上的真值，
$y_l^{(i,j)}$: 在 l-th时间步上的弱标注，来自弱标注源j)

方法优势

- 概率图模型的原则性建模，来自BERT等深度学习模型的丰富的上下文知识
- 采用Global normalization(全局优化视觉，即用$p(\mathbf{y},\mathbf{t}|\mathbf{x})=\frac{\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$)而避免了因per time-step modeling(局部优化视觉)所引起的Label Bias Problem(LBP)：
  - Global normalization (全局优化视觉)：
    - 建模：$p(\mathbf{y},\mathbf{t}|\mathbf{x})=\frac{\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}{\sum_{\mathbf{y}}\sum_{\mathbf{t}}\exp(\text{score}_\Theta(\mathbf{y},\mathbf{t},\mathbf{x}))}$
    - 模型 (即, Θ)整体性地考虑关于 “整个弱标注序列y和真值序列t依赖于句子x” 的global patterns，获得global knowledge

  - 避免LBP:
    - 因为采用了Global normalization方法（可获得灵活的路径分数，比如比如0.2, 1.5 而非0.4,0.6），而不是per time-step normalization方法(这是导致LBP的直接原因[17])，即$p(Truth_l|Truth_{l-1},Sentence;\Theta)=\frac{\exp(f_\Theta(Truth_l,Truth_{l-1},Sentence))}{\sum_{Truth_l}\exp(f_\Theta(Truth_l,Truth_{l-1},Sentence))}$
    - 监督学习算法 CRF[16] vs. MEMM[15] ≈ 弱监督学习算法 Ours 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.
[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.

其他

- BERT-CRF/CRF, Neural-Hidden-CRF 都属于概率无向图模型：
  - 随机变量的联合分布: 最大团上的势函数的得分的归一化取值 $p(X) = \frac{1}{\text{Normalization } z=\sum_X \text{SCORE}(X)} \text{SCORE}(X),\ \text{SCORE}(X) = \prod_C \phi_c(X_C)$

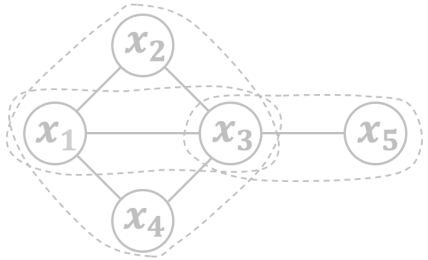


图1：一个无向图模型的例子

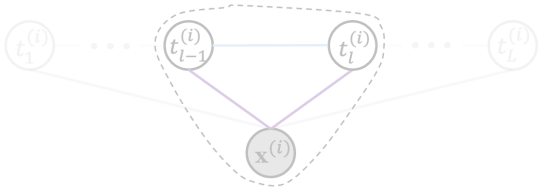


图2: BERT-CRF/CRF

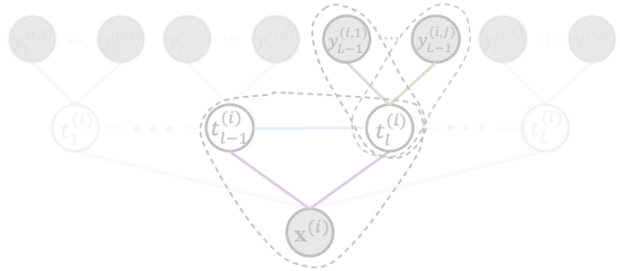


图3: Neural-Hidden-CRF

- 关于更多的公式构建与细节（模型参数的初始化、模型的计算复杂度…）， 请参考论文

实验

- 在以上内容中：为了解决“从弱监督序列标注中学习”问题，我们提出了模型Neural-Hidden-CRF；当模型完成学习后，可得到我们所关心的分类器（$\Theta_{classifier} \in \Theta_{all}$）
- 实验内容：
  - 性能分析：
    - 4个开源的弱监督数据集 （CoNLL-03 (WS), WikiGold (WS) and MIT-Restaurant (WS)来自[1], 以及CoNLL-03 (MTurk) [11]）
    - 分别调查：Prediction performance——在测试集上$\Theta_{classifier}$的预测性能； Inference performance——在有弱监督标注数据上用$\Theta_{classifier}$做真值预测的性能
  - 其他：弱监督源参数的估计结果分析、消融实验分析、其他实验分析

[1] Zhang et al. WRENCH: A Comprehensive Benchmark for Weak Supervision. NeurIPS 2021.
[11] Rodrigues et al. Deep learning from crowds. AAAI 2018.

性能分析（CoNLL-03, Wikigold, MIT-Rest. [1]）

- 分析：Prediction performance / Inference performance (超过CHMM 2.80/2.23 F1)

Paradigm	Method	Prediction on test data§			Inference on test data*			Avg.F1(P/I)
		CoNLL-03	WikiGold	MIT-Rest.	CoNLL-03	WikiGold	MIT-Rest.	
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.

性能分析（CoNLL-03 (MTurk) dataset [7]）

- 分析：Prediction performance / Inference performance

Paradigm	Method	Prediction on test data§ Precision	Recall	F1	Inference on train data* 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.

弱监督源参数分析

- 准确的弱监督源转移矩阵参数估计，良好的可解释性
  - "Real": 真实的概率混淆矩阵; "Estimated by our (1)/(2)": 通过对我们的矩阵参数做normalization计算后而得到

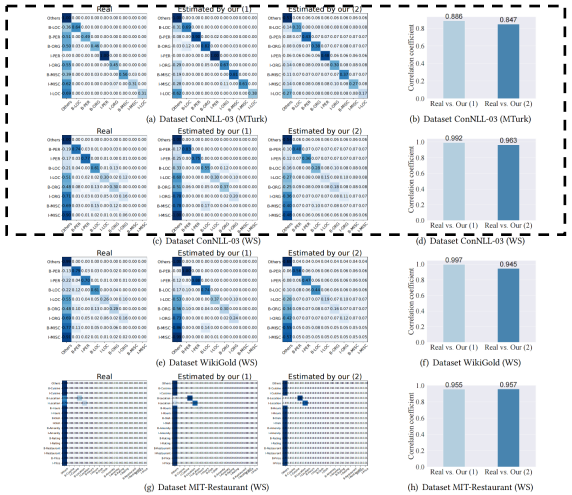


Figure: Results on four datasets.

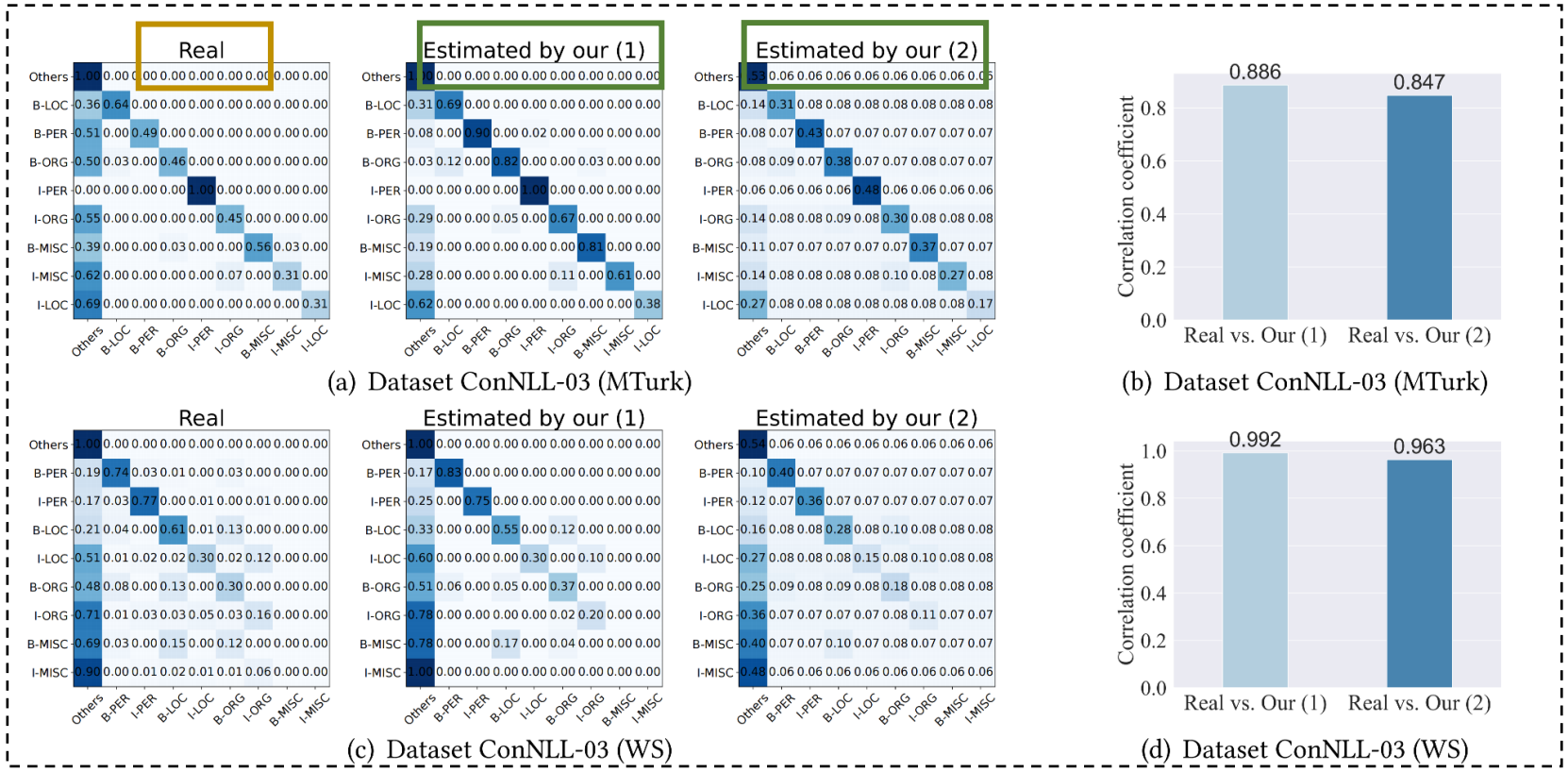
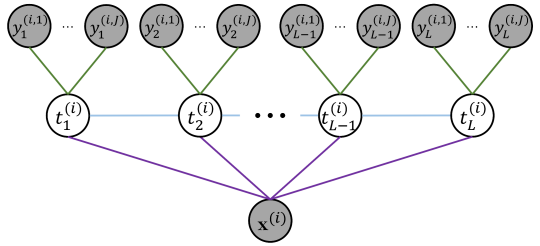


Figure: Results the the CoNLL-03 (MTurk) dataset and CoNLL-03 (WS) dataset.

消融实验

- 分别消融掉3个部分(weak source transition matrices (#1), CRF transition matrix (#2, #3), emission values (#4))
- 不合适的参数初始化 (#5, #6, #7)
- 冻结部分参数的训练 (#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.

其他实验

- Neural-Hidden-CRF配备上不同的深度学习架构（backbones）时:
  - 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).
- 更多的消融实验分析:

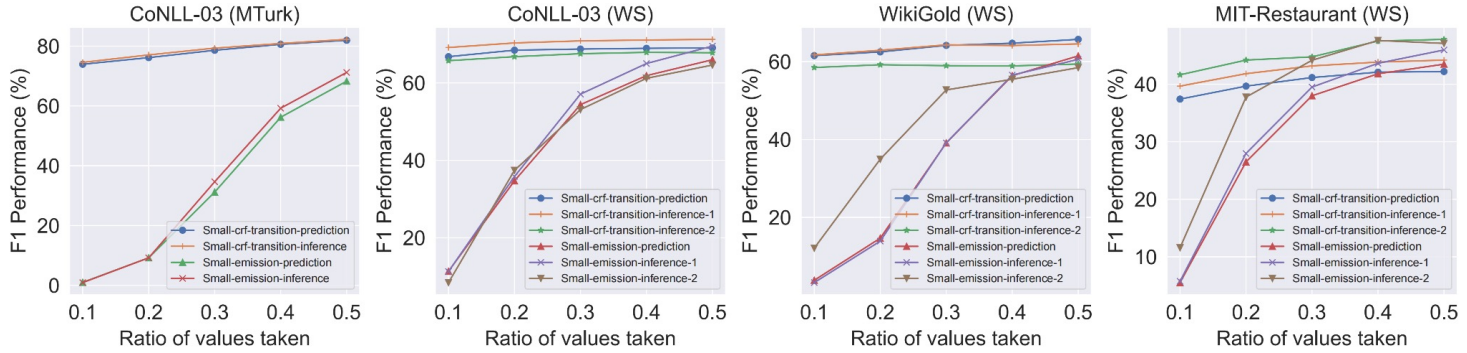


Figure A.1: Performance of more variants for supplementary ablation study. Results are averaged over 20 runs.

结论

- 从弱监督序列标注中学习（WSSL）？尝试Neural-Hidden-CRF
  - Neural-Hidden-CRF, 首个解决WSSL问题的神经化无向图模型, 它既受益于概率图模型所带来的对变量的有原则性建模，又受益于先进的深度学习模型（如BERT）所带来的丰富的上下文语义知识，同时不会引入由局部优化视角所引起的Label Bias Problem
  - Code, and more information (Slide): https://github.com/junchenzhi/Neural-Hidden-CRF