BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition

Yinghao Li, Pranav Shetty, Lucas Liu, Chao Zhang, and Le Song

Georgia Institute of Technology Mohamed bin Zayed University of Artificial Intelligence

{yinghaoli, pranav.shetty, lucasliu, chaozhang}@gatech.edu le.song@mbzuai.ac.ae

Problem setup

Weakly supervised named entity recognition (NER)

- Manually labeling NER datasets is hard and time-consuming.
- **Alternative:** automatically generate labels from *weak sources* (*e.g.*, knowledge bases, heuristic functions, *etc.*).

Multi-source weak supervision

- The annotations of one weak source are often incomplete and inaccurate.
- **Solution:** Use multiple sources to get comprehensive results.

	Rockefeller	Center	in	New	York	was...
Source 1	B-PER	O	O	B-LOC	I-LOC	O...
Source 2	B-LOC	I-LOC	O	O	B-LOC	O...
Target	B-LOC	I-LOC	O	B-LOC	I-LOC	O...

- **Input:** 1) a sequence of T tokens $w^{(1:T)}$; and 2) K sets of weak label sequences $\{x_k^{(1:T)}\}_{k=1}^K$, $x_k^{(t)} \in \mathbb{R}^L$ where L is the number of entity labels.
- **Target:** one sequence of aggregated labels $y^{(1:T)}$, $y^{(t)} \in \mathbb{R}^L$.

Conditional hidden Markov model

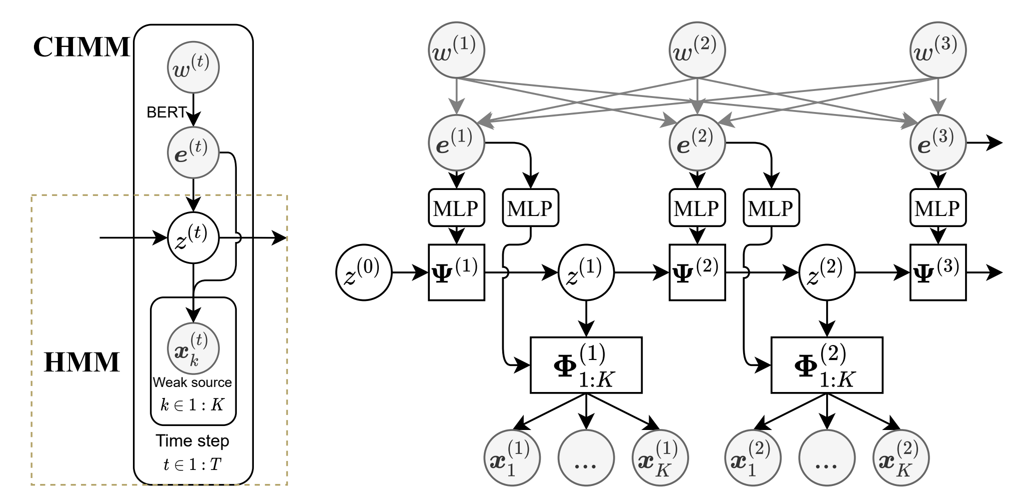
Previous approaches [5, 3] use the hidden Markov model (HMM) as the label aggregator.

- **Disadvantage:** HMM's transition and emission probabilities do not reflect input tokens' meaning and context.

The house of Barack Obama…

Ideal:	$P(\text{PER}\mid\text{others}) = 0.1$	$P(\text{PER}\mid\text{others}) = 0.8$	Different ✓
HMM:	$P(\text{PER}\mid\text{others}) = 0.2$	$P(\text{PER}\mid\text{others}) = 0.2$	Same ✗

The conditional hidden Markov model (CHMM) predicts *token-wise* transitions and emissions from the BERT token embeddings through one layer of feed-forward network.



(a) Plate notation. (b) Details.

Figure: CHMM's model architecture. z: hidden state; Ψ: transition matrix; Φ emission matrix. w represents the token and e is its BERT embedding.

Alternate-training

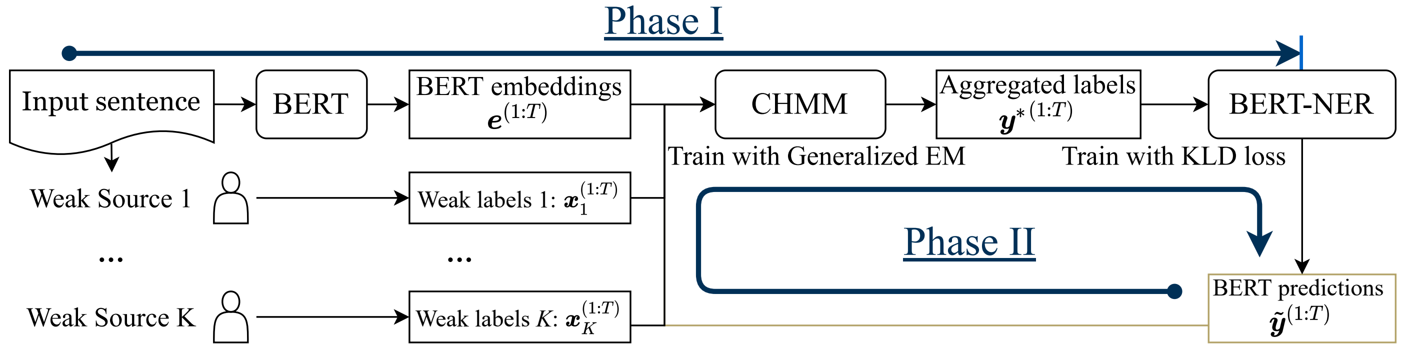
Limitation: CHMM cannot predict labels observed by no source.

		Rockefeller	Center	in	New	York	was...
Not possible!	**Source 1**	O	O	O	B-LOC	I-LOC	O...
	Source 2	O	O	O	O	B-LOC	O...
	Target	B-LOC	I-LOC	O	B-LOC	I-LOC	O...

Improvement: introduce a *supervised* BERT-NER model into the pipeline.

- BERT-NER is fine-tuned with the labels predicted by CHMM;
- BERT-NER refines the labels with the context information contained in BERT.

The alternate-training method (CHMM-ALT) trains CHMM and BERT-NER alternately in a two-phase manner.



In phase I:

- Construct weak labels $x_{1:K}^{(1:T)}$ and BERT embeddings $e^{(1:T)}$.
- Train CHMM with $x_{1:K}^{(1:T)}$ and obtain aggregated labels $y^{*(1:T)}$, $y^{*(t)} \in \mathbb{R}^L$.
- Fine-tune BERT-NER with $y^{*(1:T)}$ and KL divergence loss; get output labels $\tilde{y}^{(1:T)}$.

In phase II:

- Append BERT-NER outputs $\tilde{y}^{(1:T)}$ to weak observations: $x_{1:K+1}^{(1:T)} = \{x_{1:K}^{(1:T)}, \tilde{y}^{(1:T)}\}$.
- Train CHMM with $x_{1:K+1}^{(1:T)}$ and get its aggregated labels $y^{*(1:T)}$ as in phase I.
- Fine-tune BERT-NER from its previous checkpoint with the updated $y^{*(1:T)}$.
- Repeat the above procedure for several loops with $y^{*(1:T)}$ and $\tilde{y}^{(1:T)}$ ($x_{1:K+1}^{(1:T)}$) being alternately updated; select the best model based on the validation performance.

Experiments

	Co03	NCBI	CDR	LR
# Instance	22137	793	1500	3845
# Training	14041	593	500	2436
# Development	3250	100	100	609
# Test	3453	100	500	800
Ave# Tokens	14.5	219.8	217.7	16.4
# Entities	4	1	2	1
# Sources	13	5	8	4

Table: Dataset statistics.

Datasets:

1) CoNLL 2003 dataset of the Reuters news stories;
2) LaptopReview dataset from the custormer reviews of laptops;
3) NCBI-Disease and 4) BC5CDR datasets constructed from the biomedical science articles.

Metrics:

Entity level precision, recall and F1 scores.

Experiments

Baselines: 1) Majority Voting; 2) Snorkel [4]; 3) SwellShark [1]; 4) AutoNER [6]; 5) BOND [2]; 6) HMM [3]; 7) Linked HMM [5].

Supervised baselines: 1) BERT-NER trained with manual labels; and 2) a *best consensus* that keeps only correct annotations from each source (100% precision).

Ablation study: CHMM-iid that removes the transition dependencies of CHMM.

Main results:

Models	CoNLL 2003	NCBI-Disease	BC5CDR	LaptopReview
Supervised BERT-NER ‡ ♮	90.74 (90.37/91.10)	88.89 (87.05/90.82)	88.81 (87.12/90.57)	81.34 (82.02/80.67)
best consensus ♮	89.18 (100.0/80.47)	81.60 (100.0/68.91)	87.58 (100.0/77.89)	77.72 (100.0/63.55)
SwellShark (noun-phrase) †‡	-	67.10 (64.70/69.70)	84.23 (84.98/83.49)	-
SwellShark (hand-tuned) †‡	-	80.80 (81.60/80.10)	84.21 (86.11/82.39)	-
AutoNER †‡	67.00 (75.21/60.40)	75.52 (79.42/71.98)	82.13 (83.23/81.06)	65.44 (72.27/59.79)
Snorkel †‡	66.40 (71.40/62.10)	73.41 (71.10/76.00)	82.24 (80.23/84.35)	63.54 (64.09/63.09)
Linked HMM †‡	-	79.03 (83.46/75.05)	82.96 (82.65/83.28)	69.04 (77.74/62.11)
BOND-MV †‡ ♮	65.96 (64.22/67.82)	80.33 (84.77/76.34)	83.18 (82.90/83.49)	67.19 (68.90/65.75)
Majority Voting † ♮	58.40 (49.01/72.24)	73.94 (79.76/68.91)	80.73 (83.79/77.88)	67.92 (72.93/63.55)
HMM † ♮	68.84 (70.80/66.98)	73.06 (83.88/64.70)	80.57 (88.75/73.76)	66.96 (77.46/58.96)
CHMM-i.i.d. † ♮	68.57 (69.67/67.50)	71.69 (83.49/62.87)	79.37 (85.68/73.92)	65.89 (75.70/58.34)
CHMM † ♮	70.11 (72.98/67.47)	78.88 (**93.37**/68.28)	82.39 (**89.93**/76.02)	73.02 (**87.23**/62.79)
CHMM + BERT-NER †‡ ♮	74.30 (75.02/73.58)	82.87 (89.42/77.22)	84.33 (85.58/83.12)	69.67 (75.48/64.70)
CHMM-ALT †‡ ♮	**75.54** (**76.22**/**74.86**)	**85.02** (87.92/**82.47**)	**85.12** (84.97/**85.28**)	**76.55** (81.39/**72.32**)

Table: Metrics are presented in the "F1 (Precision/Recall)" format.

- **Observation:** CHMM has high precision; BERT-NER exchanges recall with precision.

Evaluating the alternate-training method:

Label aggregator	Co03	NCBI	CDR	Laptop	Label aggregator-ALT	Co03	NCBI	CDR	Laptop
MV † ♮	58.40	73.94	80.73	67.92	MV-ALT †‡ ♮	66.64	80.83	82.78	70.45
HMM † ♮	68.84	73.06	80.57	66.96	HMM-ALT †‡ ♮	74.04	82.99	83.34	72.90
i.i.d. † ♮	68.57	71.69	79.37	65.89	i.i.d.-ALT †‡ ♮	73.84	83.15	83.17	72.61
CHMM † ♮	70.11	78.88	82.39	73.02	CHMM-ALT †‡ ♮	75.54	85.02	85.12	76.55

Table: Alternate-training F1 scores with different label aggregators.

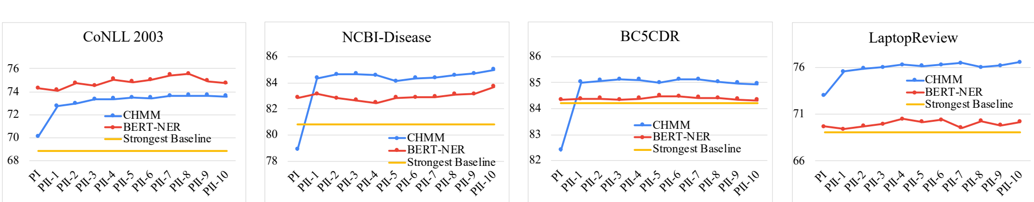


Figure: F1 score evolution through the alternate-training phases.

References

[1] J. A. Fries, S. Wu, A. Ratner, and C. Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *CoRR*, 2017.

[2] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *ACM SIGKDD*, 2020.

[3] P. Lison, J. Barnes, A. Hubin, and S. Touileb. Named entity recognition without labelled data: A weak supervision approach. In *ACL*, 2020.

[4] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282, Nov. 2017.

[5] E. Safranchik, S. Luo, and S. Bach. Weakly supervised sequence tagging from noisy rules. In *AAAI Conference*, 2020.

[6] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *EMNLP*, 2018.

https://github.com/Yinghao-Li/CHMM-ALT