Aggregating and Predicting Sequence Labels from Crowd Annotations

An T. Nguyen[1*] Byron C. Wallace[2] Jessy Li[1,3]
Ani Nenkova[3] Matthew Lease[1]

[1]University of Texas at Austin

[2] Northeastern University

[3] University of Pennsylvania

ACL 2017

[*]Presenter

Problem: Sequence Labeling with Crowd Labels

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc
W_2	Org	Per	Per	O	O	Loc

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc
W_2	Org	Per	Per	O	O	Loc
W_3	Org	O	Per	O	O	Loc

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc
W_2	Org	Per	Per	O	O	Loc
W_3	Org	O	Per	O	O	Loc

Two tasks:

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc
W_2	Org	Per	Per	O	O	Loc
W_3	Org	O	Per	O	O	Loc

Two tasks:

- Aggregation: Given $(X, W_{1,2,3})$, Estimate Y

Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

X	U.N.	official	Ekeus	heads	for	Baghdad
Y	Org	O	Per	O	O	Loc
W_1	Org	O	Org	O	O	Loc
W_2	Org	Per	Per	O	O	Loc
W_3	Org	O	Per	O	O	Loc

Two tasks:

- Aggregation: Given $(X, W_{1,2,3})$, Estimate Y
- Prediction: Given train data $(X, W_{1,2,3})$, Predict Y_{test} for X_{test}

Our work

Contribution: Two **Joint models** of sequences and crowd.

Our work

Contribution: Two **Joint models** of sequences and crowd.

1. Aggregation.
   - Hidden Markov Models (HMMs) $+$
     Crowd Confusion Matrices.

Our work

Contribution: Two **Joint models** of sequences and crowd.

1. Aggregation.
   - Hidden Markov Models (HMMs) + Crowd Confusion Matrices.
2. Prediction.
   - Long Short Term memory (LSTM) + Crowd Embedding Vectors.

Our work

Contribution: Two **Joint models** of sequences and crowd.

1. Aggregation.
   - Hidden Markov Models (HMMs) + Crowd Confusion Matrices.
2. Prediction.
   - Long Short Term memory (LSTM) + Crowd Embedding Vectors.

Evaluation:

- News NER + Biomedical IE.
- A range of baselines.

Our work

Contribution: Two **Joint models** of sequences and crowd.

1. Aggregation.
   - Hidden Markov Models (HMMs) + Crowd Confusion Matrices.
2. Prediction.
   - Long Short Term memory (LSTM) + Crowd Embedding Vectors.

Evaluation:

- News NER + Biomedical IE.
- A range of baselines.

Code + Data on Github.

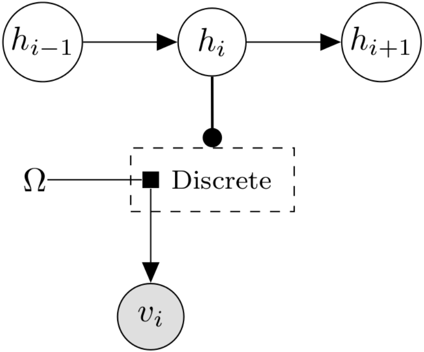
HMM-Crowd

(for task 1 - aggregation)

HMM (position i):

$$h_{i+1} | h_i \sim \mathrm{Discrete}(\boldsymbol{\tau}_{h_i})$$

$$v_i | h_i \sim \mathrm{Discrete}(\boldsymbol{\Omega}_{h_i})$$



HMM-Crowd

(for task 1 - aggregation)

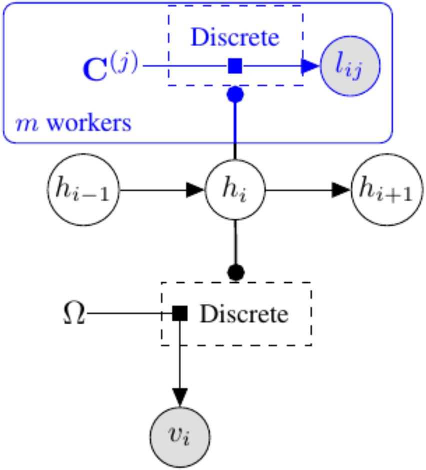
HMM (position i):

$$h_{i+1} | h_i \sim \mathrm{Discrete}(\boldsymbol{\tau}_{h_i})$$

$$v_i | h_i \sim \mathrm{Discrete}(\boldsymbol{\Omega}_{h_i})$$

Crowd model (worker j):

$$l_{ij} | h_i \sim \mathrm{Discrete}(\mathbf{C}^{(j)}_{h_i})$$



HMM-Crowd

(for task 1 - aggregation)

HMM (position i):

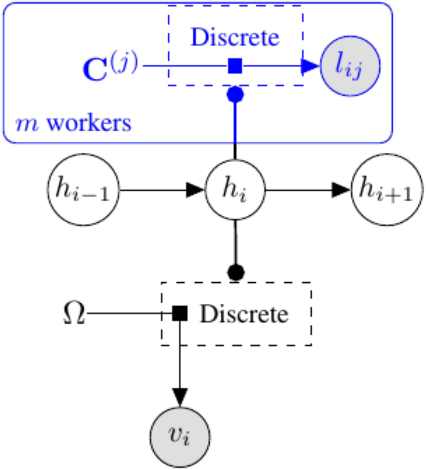
$$h_{i+1}|h_i \sim \text{Discrete}(\boldsymbol{\tau}_{h_i})$$

$$v_i|h_i \sim \text{Discrete}(\boldsymbol{\Omega}_{h_i})$$

Crowd model (worker j):

$$l_{ij}|h_i \sim \text{Discrete}(\mathbf{C}^{(j)}_{h_i})$$

$\mathbf{C}^{(j)}$: confusion matrix for j



HMM-Crowd: Parameter Learning

Expectation Maximization (EM) algorithm:

HMM-Crowd: Parameter Learning

Expectation Maximization (EM) algorithm:

E-step

- Estimate posterior $p(h)$
- Extend Forward-Backward algorithm.

HMM-Crowd: Parameter Learning

Expectation Maximization (EM) algorithm:

E-step

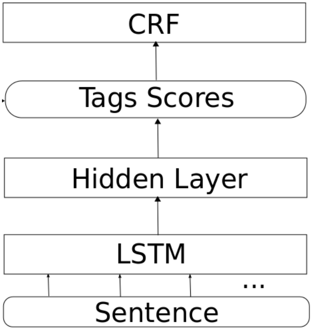
- Estimate posterior $p(h)$
- Extend Forward-Backward algorithm.

M-step:

- Estimate parameters τ, Ω, C
- Variational Bayes estimate.

LSTM for NER

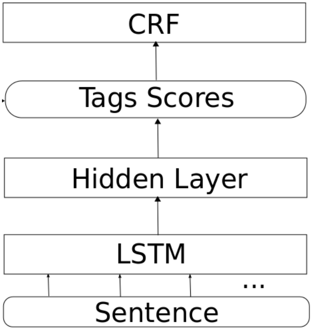
(Lample et al. 2016)



LSTM for NER

(Lample et al. 2016)

LSTM: word rep. $\rightarrow$ sent. rep.

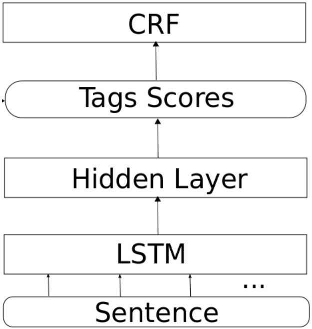


LSTM for NER

(Lample et al. 2016)

LSTM: word rep. $\rightarrow$ sent. rep.

Hidden Layer: fully connected.



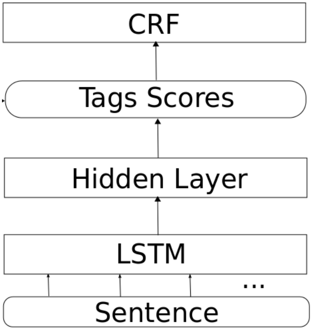
LSTM for NER

(Lample et al. 2016)

LSTM: word rep. $\rightarrow$ sent. rep.

Hidden Layer: fully connected.

Tags Scores: $\sim$ prob. each label for each word.



LSTM for NER

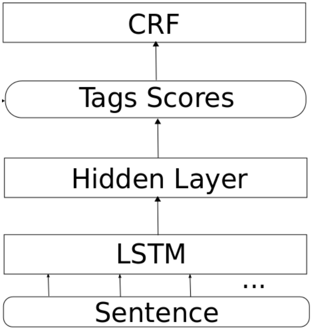
(Lample et al. 2016)

LSTM: word rep. $\rightarrow$ sent. rep.

Hidden Layer: fully connected.

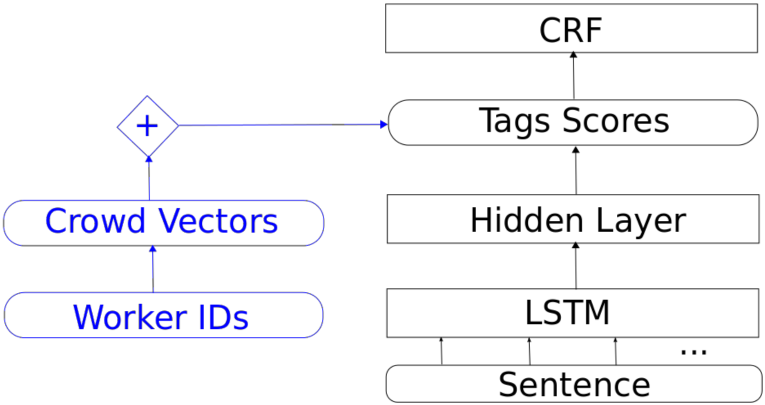
Tags Scores: $\sim$ prob. each label for each word.

CRF: word prediction $\rightarrow$ sent. prediction.



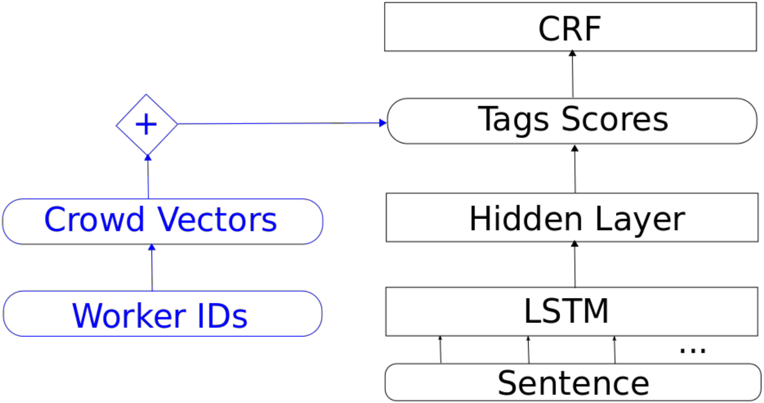
LSTM-Crowd

(for task 2 - prediction)



LSTM-Crowd

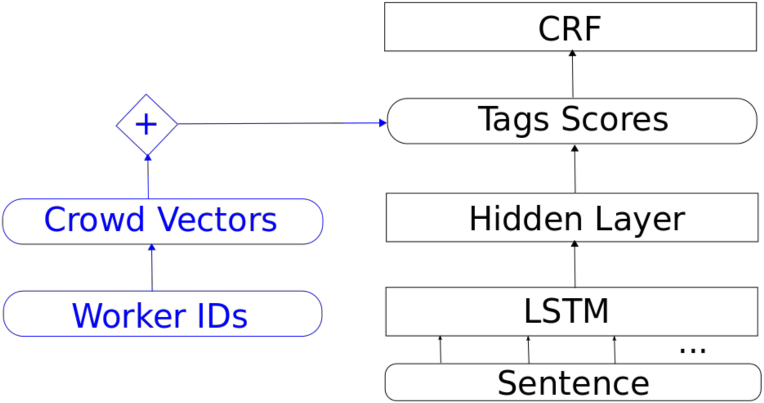
(for task 2 - prediction)



- vectors represented noise by worker.

LSTM-Crowd

(for task 2 - prediction)



- vectors represented noise by worker.
- v(good worker) ≈ 0

Data

Dataset	Application	Documents	Gold Labels	Crowd Labels
CoNLL'03	NER	1393	All	400
Medical	IE	5000	200	All

Evaluation: Task 1 - aggregation

Baselines:

1. Non-sequential:
   - Majority Voting
   - Dawid & Skene (1979)
   - MACE (Hovy et al. 2013)

Evaluation: Task 1 - aggregation

Baselines:

1. Non-sequential:
   - Majority Voting
   - Dawid & Skene (1979)
   - MACE (Hovy et al. 2013)
2. Sequential:
   - CRF-MA (Rodrigues et al. 2014)

Results: NER task 1 - aggregation

Method	F1
Majority Vote	65.71

Results: NER task 1 - aggregation

Method	F1
Majority Vote	65.71
MACE (Hovy et al. 2013)	67.37

Results: NER task 1 - aggregation

Method	F1
Majority Vote	65.71
MACE (Hovy et al. 2013)	67.37
Dawid-Skene (DS)	71.39

Results: NER task 1 - aggregation

Method	F1
Majority Vote	65.71
MACE (Hovy et al. 2013)	67.37
Dawid-Skene (DS)	71.39
CRF-MA (Rodrigues et al. 2014)	62.53

Results: NER task 1 - aggregation

Method	F1
Majority Vote	65.71
MACE (Hovy et al. 2013)	67.37
Dawid-Skene (DS)	71.39
CRF-MA (Rodrigues et al. 2014)	62.53
HMM-Crowd	**74.76**

Evaluation: Task 2 - prediction

Baselines:

1. Aggregate then train:
   - Majority Vote then CRF
   - Dawid-Skene then LSTM

Evaluation: Task 2 - prediction

Baselines:

1. Aggregate then train:
   - Majority Vote then CRF
   - Dawid-Skene then LSTM
2. Train directly on crowd labels:
   - CRF-MA (Rodrigues et al. 2014)
   - LSTM (original, Lample et al. 2016)

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60
LSTM (Lample et al. 2016)	67.73

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60
LSTM (Lample et al. 2016)	67.73
Dawid-Skene then LSTM	66.27

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60
LSTM (Lample et al. 2016)	67.73
Dawid-Skene then LSTM	66.27
LSTM-Crowd	70.82

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60
LSTM (Lample et al. 2016)	67.73
Dawid-Skene then LSTM	66.27
LSTM-Crowd	70.82
HMM-Crowd then LSTM	**70.87**

Results: NER task 2 - prediction

Method	F1
Majority Vote then CRF	58.20
CRF-MA (Rodrigues et al. 2014)	62.60
LSTM (Lample et al. 2016)	67.73
Dawid-Skene then LSTM	66.27
LSTM-Crowd	70.82
HMM-Crowd then LSTM	**70.87**
LSTM on Gold Labels (upper-bound)	84.22

Conclusion

- Joint models of sequences and crowd labels.

Conclusion

- Joint models of sequences and crowd labels.
- HMMs good for aggregation, ...
- ... LSTMs good for prediction.

Conclusion

- Joint models of sequences and crowd labels.
- HMMs good for aggregation, ...
- ... LSTMs good for prediction.

Paper:

- Alternative LSTM-Crowd model.
- Results for Biomedical IE.

Conclusion

- Joint models of sequences and crowd labels.
- HMMs good for aggregation, ...
- ... LSTMs good for prediction.

Paper:

- Alternative LSTM-Crowd model.
- Results for Biomedical IE.

Conclusion

- Joint models of sequences and crowd labels.
- HMMs good for aggregation, ...
- ... LSTMs good for prediction.

Paper:

- Alternative LSTM-Crowd model.
- Results for Biomedical IE.

Acknowledgment: Reviewers, Workers, NSF & NIH.

Conclusion

- Joint models of sequences and crowd labels.
- HMMs good for aggregation, ...
- ... LSTMs good for prediction.

Paper:

- Alternative LSTM-Crowd model.
- Results for Biomedical IE.

Acknowledgment: Reviewers, Workers, NSF & NIH.

Questions?