Aggregating and Predicting Sequence Labels from CrowdAnnotations

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

Example: Named Entity Recognition.

```
U.N. official Ekeus heads for Baghdad

1: Org
2: Org Per
3: Org O

Two tasks:
- Aggregation: Given (X, W_1, W_2, W_3), Estimate Y
- Prediction: Given train data (X, W_1, W_2, W_3), Predict Y_{test} for X_{test}
```
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N. official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_1$</td>
<td>Org</td>
<td>O</td>
<td>Org</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_1$</td>
<td>Org</td>
<td>O</td>
<td>Org</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_2$</td>
<td>Org</td>
<td>Per</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₁</td>
<td>Org</td>
<td>O</td>
<td><strong>Org</strong></td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₂</td>
<td>Org</td>
<td><strong>Per</strong></td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₃</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_1$</td>
<td>Org</td>
<td>O</td>
<td><strong>Org</strong></td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_2$</td>
<td>Org</td>
<td></td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_3$</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>

Two tasks:
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₁</td>
<td>Org</td>
<td>O</td>
<td>Org</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₂</td>
<td>Org</td>
<td>Per</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>W₃</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>

Two tasks:

- Aggregation: Given \((X, W_{1,2,3})\), Estimate \(Y\)
Problem: Sequence Labeling with Crowd Labels

Example: Named Entity Recognition.

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>U.N.</th>
<th>official</th>
<th>Ekeus</th>
<th>heads</th>
<th>for</th>
<th>Baghdad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_1$</td>
<td>Org</td>
<td>O</td>
<td><strong>Org</strong></td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_2$</td>
<td>Org</td>
<td>Per</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
<tr>
<td>$W_3$</td>
<td>Org</td>
<td>O</td>
<td>Per</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>Loc</td>
</tr>
</tbody>
</table>

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.

2. Prediction.

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

Code + Data on Github.
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 $i$):

$$h_{i+1}|h_i \sim \text{Discrete}(\tau_{h_i})$$

$$v_i|h_i \sim \text{Discrete}(\Omega_{h_i})$$
HMM-Crowd
(for task 1 - aggregation)

HMM (position $i$):

$h_{i+1}|h_i \sim \text{Discrete}(\tau_{h_i})$

$v_i|h_i \sim \text{Discrete}(\Omega_{h_i})$

Crowd model (worker $j$):

$l_{ij}|h_i \sim \text{Discrete}(C_{h_i}^{(j)})$
HMM-Crowd
(for task 1 - aggregation)

HMM (position $i$):

$$h_{i+1}|h_i \sim \text{Discrete}(\tau_{h_i})$$

$$v_i|h_i \sim \text{Discrete}(\Omega_{h_i})$$

Crowd model (worker $j$):

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

$C^{(j)}$: confusion matrix for $j$
HMM-Crowd: Parameter Learning

Expectation Maximization (EM) algorithm:

1. **E-step**: Estimate posterior $p(h)$
2. **M-step**: Estimate parameters $\tau$, $\Omega$, $C$
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 $\tau, \Omega, C$
- 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. → 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. → sent. rep.

Hidden Layer: fully connected.

Tags Scores: \( \sim \) prob. each label for each word.

CRF: word prediction → 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.
- $v(\text{good worker}) \approx 0$
## Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Application</th>
<th>Documents</th>
<th>Gold Labels</th>
<th>Crowd Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL’03</td>
<td>NER</td>
<td>1393</td>
<td>All</td>
<td>400</td>
</tr>
<tr>
<td>Medical</td>
<td>IE</td>
<td>5000</td>
<td>200</td>
<td>All</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote</td>
<td>65.71</td>
</tr>
</tbody>
</table>
### Results: NER task 1 - aggregation

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote</td>
<td>65.71</td>
</tr>
<tr>
<td>MACE (Hovy et al. 2013)</td>
<td>67.37</td>
</tr>
</tbody>
</table>
## Results: NER task 1 - aggregation

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote</td>
<td>65.71</td>
</tr>
<tr>
<td>MACE (Hovy et al. 2013)</td>
<td>67.37</td>
</tr>
<tr>
<td>Dawid-Skene (DS)</td>
<td>71.39</td>
</tr>
</tbody>
</table>
Results: NER task 1 - aggregation

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote</td>
<td>65.71</td>
</tr>
<tr>
<td>MACE (Hovy et al. 2013)</td>
<td>67.37</td>
</tr>
<tr>
<td>Dawid-Skene (DS)</td>
<td>71.39</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.53</td>
</tr>
</tbody>
</table>
Results: NER task 1 - aggregation

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote</td>
<td>65.71</td>
</tr>
<tr>
<td>MACE (Hovy et al. 2013)</td>
<td>67.37</td>
</tr>
<tr>
<td>Dawid-Skene (DS)</td>
<td>71.39</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.53</td>
</tr>
<tr>
<td>HMM-Crowd</td>
<td>74.76</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
</tbody>
</table>
## Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
</tbody>
</table>
## Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
<tr>
<td>LSTM (Lample et al. 2016)</td>
<td>67.73</td>
</tr>
</tbody>
</table>
## Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
<tr>
<td>LSTM (Lample et al. 2016)</td>
<td>67.73</td>
</tr>
<tr>
<td>Dawid-Skene then LSTM</td>
<td>66.27</td>
</tr>
</tbody>
</table>
## Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
<tr>
<td>LSTM (Lample et al. 2016)</td>
<td>67.73</td>
</tr>
<tr>
<td>Dawid-Skene then LSTM</td>
<td>66.27</td>
</tr>
<tr>
<td>LSTM-Crowd</td>
<td>70.82</td>
</tr>
</tbody>
</table>
Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
<tr>
<td>LSTM (Lample et al. 2016)</td>
<td>67.73</td>
</tr>
<tr>
<td>Dawid-Skene then LSTM</td>
<td>66.27</td>
</tr>
<tr>
<td>LSTM-Crowd</td>
<td>70.82</td>
</tr>
<tr>
<td>HMM-Crowd then LSTM</td>
<td>70.87</td>
</tr>
</tbody>
</table>
## Results: NER task 2 - prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Vote then CRF</td>
<td>58.20</td>
</tr>
<tr>
<td>CRF-MA (Rodrigues et al. 2014)</td>
<td>62.60</td>
</tr>
<tr>
<td>LSTM (Lample et al. 2016)</td>
<td>67.73</td>
</tr>
<tr>
<td>Dawid-Skene then LSTM</td>
<td>66.27</td>
</tr>
<tr>
<td>LSTM-Crowd</td>
<td>70.82</td>
</tr>
<tr>
<td>HMM-Crowd then LSTM</td>
<td><strong>70.87</strong></td>
</tr>
<tr>
<td><em>LSTM on Gold Labels (upper-bound)</em></td>
<td>84.22</td>
</tr>
</tbody>
</table>
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?