BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision

Chen Liang*, Yue Yu*, Haoming Jiang*, Siawpeng Er, Ruijia Wang, Tuo Zhao, Chao Zhang

Georgia Institute of Technology

* Equal contributions
When Sebastian Thrun started at Google in 2007, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode earlier this week.
When Sebastian Thrun \textcolor{green}{PERSON} started at \textcolor{blue}{Google} \textcolor{green}{ORG} in \textcolor{teal}{2007} \textcolor{green}{DATE}, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American \textcolor{pink}{NORP} car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun \textcolor{green}{PERSON}, now the co-founder and CEO of online higher education startup Udacity, in an interview with \textcolor{cyan}{Recode} \textcolor{blue}{ORG} earlier this week \textcolor{teal}{DATE}. 
When **Sebastian Thrun** started at **Google** in **2007**, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said **Thrun**, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** earlier this week.
When **Sebastian Thrun** started at **Google** in **2007**, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn’t worth talking to,” said **Thrun**, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** earlier this week.
Sounds like a very confusing comment then
Distant Supervision
Distant Supervision
POS Tagging: ... NNP NNP NNP ... 
Sentence: ... Liverpool Football Club ... 
Potential Entity
POS Tagging: ... NNP NNP NNP ...
Sentence: ... Liverpool Football Club ...
Potential Entity
Liverpool: Q16977
ID in WikiData
Query ID
WikiData
SELECT ?item WHERE {
  FILTER ( ?item in (Q1733492) )
}

SELECT ?item WHERE {
  FILTER ( ?item in (Q215627) )
}
POS Tagging: NNP NNP NNP ...
Sentence: Liverpool Football Club ...

Potential Entity

Location: Q1733492
Entity Type ID in WikiData

Organization: Q215627

SELECT ?item WHERE {
  FILTER ( ?item in (Q1733492) )
}

SELECT ?item WHERE {
  FILTER ( ?item in (Q215627) )
}

ID in WikiData

WIKIDATA

Matched

Matched!

Query ID

Noisy Annotation!
SELECT ?item WHERE {
FILTER (?item in (Q1733492))}

SELECT ?item WHERE {
FILTER (?item in (Q215627))}
Open Domain NER with Distant Supervision

Matching Performance on **Open-Domain** vs. **Single-Domain** NER Data

- F1-Scores
- Precision
- Recall

- Open-Domain: CoNLL03
- Medical Domain: BC5CDR
BOND: BERT-Assisted Open Domain NER with Distant Supervision
BOND: BERT-Assisted Open Domain NER with Distant Supervision

- Leverage power of **pre-trained language models** (e.g. BERT, RoBERTa)
BOND: BERT-Assisted Open Domain NER with Distant Supervision

- Leverage power of pre-trained language models (e.g. RoBERTa)
- Two-stage self-training framework
...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
Stage I: Pseudo Label Generation

Lynn Reaser: **PER**
Barnett Banks Inc: **ORG**
Jacksonville: **ORG**

Knowledge Bases
Multi-source Gazetteer

Distant Labels

Classification Head
RoBERTa
Pretrained

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Knowledge Bases

Multi-source Gazetteer

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Stage I: Pseudo Label Generation

Lynn Reaser: PER
Barnett Banks Inc: ORG
Jacksonville: ORG

Classification Head

RoBERTa

Pretrained

Distant Labels

Pseudo Labels
Stage I: Pseudo Label Generation

Knowledge Bases
Multi-source Gazetteer

Lynn Reaser: PER
Barnett Banks Inc: ORG
Jacksonville: ORG

Knowledge Bases
Multi-source Gazetteer

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Stage II: Self-Training

Pretrained RoBERTa
Classification Head

Student Model

Teacher Model

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
Stage I: Pseudo Label Generation

- Knowledge Bases
- Multi-source Gazetteer
- Freebase

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Stage II: Self-Training

Classification Head

RoBERTa

Pretrained

Initialize

Teacher Model

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Student Model

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

O B-PER  I-PER  O B-ORG  I-ORG  I-ORG  O B-LOC
Lynn Reaser of Barnett Banks Inc. in Jacksonville...
...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

**Stage I: Pseudo Label Generation**

- Knowledge Bases
- Multi-source Gazetteer
- Freebase

Lynn Reaser: PER
Barnett Banks Inc: ORG
Jacksonville: ORG

- Classification Head
- RoBERTa: Pretrained

- Stage II: Self-Training

- Classification Head
- RoBERTa: Student Model

- Refined Pseudo Labels
- Periodic Update

- Teacher Model
Stage I: Pseudo Label Generation

Classification Head

RoBERTa

Pretrained

Distant Labels

Stage II: Self-Training

Classification Head

RoBERTa

Student Model

Initialize

Periodic Update

Teacher Model

I-O B-PER I-PER O B-ORG

I-ORG I-ORG O B-LOC

Refined Pseudo Labels

Pseudo Labels

Distant Labels

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
Stage I: BERT-Assisted Distantly Supervised Learning

...economist Lynn Reaser of Barette ...

(Open Domain Distantly Labeled Data)
Stage I: BERT-Assisted Distantly Supervised Learning

- RoBERTa Masked LM Classification Head
- Masked Language Model

- RobBERTa
- NER Classification Head
- Pretrained

Transfer

...economist MASK MASK of Barette ...
(Open Domain Unlabeled Data)

...economist Lynn Reaser of Barette ...
(Open Domain Distantly Labeled Data)
Stage I: Early Stopping

BERT Pre-trained Embedding

Overfitting

Early Stopping

Sample Output

Distant True

Distant False

Model Output

Ideal

Model

Embedding Space

Positive

Negative

Positive

Negative

Positive

Negative

Positive

Negative
Stage I: Pseudo Label Generation

Stage II: Self-Training

Lynn Reaser: PER
Barnett Banks Inc: ORG
Jacksonville: ORG

Knowledge Bases
Multi-source Gazetteer

RoBERTa: Pretrained
Classification Head

Pseudo Labels
Distant Labels

Stage I: Pseudo Label Generation

Stage II: Self-Training

Pretrained
Classification Head

Teacher Model
Student Model

Periodic Update
Refined Pseudo Labels

Distant Labels
Pseudo Labels

...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...
...economist Lynn Reaser of Barnett Banks Inc. in Jacksonville...

Initial Setup

 initialization

...
Stage II: Teacher-Student Framework

- Initialize the teacher model $f(\cdot; \theta_{tea})$ and the student model $f(\cdot; \theta_{stu})$ with the early stopped model $f(\cdot; \hat{\theta})$ obtained in Stage I

$$
\theta_{tea}^{(0)} = \theta_{stu}^{(0)} = \hat{\theta}
$$
Stage II: Teacher-Student Framework

- Initialize the teacher model $f(\cdot; \theta_{\text{tea}})$ and the student model $f(\cdot; \theta_{\text{stu}})$ with the early stopped model $f(\cdot; \hat{\theta})$ obtained in Stage I
  \[
  \theta^{(0)}_{\text{tea}} = \theta^{(0)}_{\text{stu}} = \hat{\theta}
  \]

- At the $t$-th iteration
  - The teacher model generates pseudo labeled data $\{X_m, \tilde{Y}_m\}_{m=1}^M$
  - The teacher model generates pseudo labeled data $\tilde{y}_{m,n}^{(t)} = \arg\max_c f_{n,c}(X_m; \theta_{\text{tea}}^{(t)})$
Stage II: Teacher-Student Framework

• Initialize the teacher model $f(\cdot; \theta_{tea})$ and the student model $f(\cdot; \theta_{stu})$ with the early stopped model $f(\cdot; \hat{\theta})$ obtained in Stage I

$$
\theta_{tea}^{(0)} = \theta_{stu}^{(0)} = \hat{\theta}
$$

• At the $t$-th iteration
  
  • The teacher model generates pseudo labeled data \( \{X_m, \tilde{Y}_m\}_{m=1}^M \)
  
  $$
  \tilde{y}_{m,n}^{(t)} = \arg\max_c f_{n,c}(X_m; \theta_{tea}^{(t)})
  $$

  • The student model fits these pseudo-labels by solving

  $$
  \theta_{stu}^{(t)} = \arg\min_{\theta} \frac{1}{M} \sum_{m=1}^M \ell(\tilde{Y}_m^{(t)}, f(X_m; \theta))
  $$
**Stage II: Teacher-Student Framework**

- Initialize the teacher model \( f(\cdot; \theta_{tea}) \) and the student model \( f(\cdot; \theta_{stu}) \) with the early stopped model \( f(\cdot; \hat{\theta}) \) obtained in Stage I

\[
\theta^{(0)}_{tea} = \theta^{(0)}_{stu} = \hat{\theta}
\]

- At the \( t \)-th iteration
  - The teacher model generates pseudo labeled data \( \{X_m, \tilde{Y}_m\}_{m=1}^{M} \)
    \[
    \tilde{y}^{(t)}_{m,n} = \arg\max_c f_{n,c}(X_m; \theta^{(t)}_{tea})
    \]
  - The student model fits these pseudo-labels by solving
    \[
    \theta^{(t)}_{stu} = \arg\min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \ell(\tilde{Y}_m^{(t)}, f(X_m; \theta))
    \]
  - Update the teacher model and the student model by
    \[
    \theta^{(t+1)}_{tea} = \theta^{(t+1)}_{stu} = \hat{\theta}^{(t)}_{stu}
    \]
Stage II: Teacher-Student Framework

- Stage II: With Distant Labels
- Stage II: With Pseudo-Labels

Model Output:
- Ideal
- Stage I
- Stage II

Positive
Negative
Embedding Space
**Stage II: Soft Labels w/ Confidence Re-weighting**

- At the $t$-th iteration
  - Denote the output probability simplex over $C$ classes as $[f_{n,c}(X_m; \theta)]_{c=1}^C$
Stage II: Soft Labels w/ Confidence Re-weighting

• At the $t$-th iteration
  • Denote the output probability simplex over $C$ classes as $[f_{n,c}(X_m; \theta)]_{c=1}^C$
  • The teacher model compute the unnormalized frequency of the tokens belonging to the $c$-th class as

$$p_c = \sum_{m=1}^M \sum_{n=1}^N f_{n,c}(X_m; \theta_{tea}^{(t)})$$
Stage II: Soft Labels w/ Confidence Re-weighting

• At the $t$-th iteration
  • Denote the output probability simplex over $C$ classes as $[f_{n,c}(X_m; \theta)]_{c=1}^{C}$
  • The teacher model computes the unnormalized frequency of the tokens belonging to the $c$-th class as
    \[ p_c = \sum_{m=1}^{M} \sum_{n=1}^{N} f_{n,c}(X_m; \theta_{tea}^{(t)}) \]
  • The teacher model generates soft labels $\left\{ S_m^{(t)} = \left[ s_{m,n}^{(t)} \right]_{n=1}^{N} \right\}_{m=1}^{M}$ by
    \[ s_{m,n}^{(t)} = \left[ s_{m,n,c}^{(t)} \right]_{c=1}^{C} = \left[ \frac{f_{n,c}^2(X_m; \theta_{tea}^{(t)})/p_c}{\sum_{c'=1}^{C} f_{n,c'}^2(X_m; \theta_{tea}^{(t)})/p_{c'}} \right]_{c=1}^{C} \]
Stage II: Soft Labels w/ Confidence Re-weighting

- At the $t$-th iteration
  - Denote the output probability simplex over $C$ classes as $[f_{n,c}(X_m; \theta)]_{c=1}^C$
  - The teacher model compute the unnormalized frequency of the tokens belonging to the $c$-th class as
    \[
    p_c = \sum_{m=1}^{M} \sum_{n=1}^{N} f_{n,c}(X_m; \theta_{tea}^{(t)})
    \]
  - The teacher model generates soft labels $S^{(t)}_m = \left[ s^{(t)}_{m,n,c} \right]_{c=1}^C$ by
    \[
    s^{(t)}_{m,n,c} = \frac{f_{n,c}^2(X_m; \theta_{tea}^{(t)}) / p_c}{\sum_{c'=1}^{C} f_{n,c'}^2(X_m; \theta_{tea}^{(t)}) / p_{c'}}
    \]
  - The student model fit the soft labels by solving
    \[
    \theta_{stu}^{(t)} = \arg\min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \ell_{KL}(S^{(t)}_m, f(X_m; \theta)),
    \]
    where \[
    \ell_{KL}(S^{(t)}_m, f(X_m; \theta)) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} -s^{(t)}_{m,n,c} \log f_{n,c}(X_m; \theta)
    \]
Stage II: High-Confidence Selection

• At the $t$-th iteration
  • Select a set of high confidence tokens from the $m$-th sentence by
    \[ H_m^{(t)} = \left\{ n : \max_c s_{m,n,c}^{(t)} > \epsilon \right\}, \]
    where $\epsilon \in (0,1)$
Stage II: High-Confidence Selection

• At the $t$-th iteration
  • Select a set of high confidence tokens from the $m$-th sentence by
    \[ H_m^{(t)} = \{ n : \max_c s_{m,n,c}^{(t)} > \epsilon \}, \]
    where $\epsilon \in (0,1)$

• The student model fit the high-confidence labels of the selected tokens by solving
  \[
  \theta_{stu}^{(t)} = \arg\min_{\theta} \frac{1}{M |H_{m}^{(t)}|} \sum_{m=1}^{M} \sum_{n \in H_{m}^{(t)}} -s_{m,n,c}^{(t)} \log f_{n,c}(X_{m}; \theta)
  \]
### Experiment: Main Result

#### Table 2: Main Results on Testing Set: $F_1$ Score (Precision/Recall) (in %)

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Tweet</th>
<th>OntoNote5.0</th>
<th>Webpage</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity Types</strong></td>
<td>4</td>
<td>10</td>
<td>18</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>KB Matching</td>
<td>71.40(81.13/63.75)</td>
<td>35.83(40.34/32.22)</td>
<td>59.51(63.86/55.71)</td>
<td>52.45(62.59/45.14)</td>
<td>47.76(47.90/47.63)</td>
</tr>
<tr>
<td><strong>Fully-Supervised</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>90.11(89.14/91.10)</td>
<td>52.19(51.76/52.63)</td>
<td>86.20(84.59/87.88)</td>
<td>72.39(66.29/79.73)</td>
<td>86.43(85.33/87.56)</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>91.21(91.35/91.06)</td>
<td>52.18(60.01/46.16)</td>
<td>86.17(85.99/86.36)</td>
<td>52.34(50.07/54.76)</td>
<td>54.90(55.40/54.30)</td>
</tr>
<tr>
<td><strong>Baseline</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>59.50(75.50/49.10)</td>
<td>21.77(46.91/14.18)</td>
<td>66.41(68.44/64.50)</td>
<td>43.34(58.05/34.59)</td>
<td>42.92(47.55/39.11)</td>
</tr>
<tr>
<td>AutoNER</td>
<td>67.00(75.21/60.40)</td>
<td>26.10(43.26/18.69)</td>
<td>67.18(64.63/69.95)</td>
<td>51.39(48.82/54.23)</td>
<td>47.54(43.54/52.35)</td>
</tr>
<tr>
<td>LRNT</td>
<td>69.74(79.91/61.87)</td>
<td>23.84(46.94/15.98)</td>
<td>67.69(67.36/68.02)</td>
<td>47.74(46.70/48.83)</td>
<td>46.21(45.60/46.84)</td>
</tr>
<tr>
<td><strong>Other Baseline</strong> (Reported Results)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KALM †</td>
<td>76.00(—/—)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ConNET ‡</td>
<td>75.57(84.11/68.61)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Our BOND Framework</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>46.61(53.11/41.52)</td>
<td>68.11(66.71/69.56)</td>
<td>59.11(60.14/58.11)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>BOND</td>
<td>81.48(82.05/80.92)</td>
<td>48.01(53.16/43.76)</td>
<td>68.35(67.14/69.61)</td>
<td>65.74(67.37/64.19)</td>
<td>60.07(53.44/68.58)</td>
</tr>
</tbody>
</table>
## Experiment: Main Result

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Tweet</th>
<th>OntoNote5.0</th>
<th>Webpage</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity Types</strong></td>
<td>4</td>
<td>10</td>
<td>18</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>KB Matching</td>
<td>71.40(81.13/63.75)</td>
<td>35.83(40.34/32.22)</td>
<td>59.51(63.86/55.71)</td>
<td>52.45(62.59/45.14)</td>
<td>47.76(47.90/47.63)</td>
</tr>
<tr>
<td><strong>Fully-Supervised</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>90.11(89.14/91.10)</td>
<td>52.19(51.76/52.63)</td>
<td>86.20(84.59/87.88)</td>
<td>72.39(66.29/79.73)</td>
<td>86.43(85.33/87.56)</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>91.21(91.35/91.06)</td>
<td>52.18(60.01/46.16)</td>
<td>86.17(85.99/86.36)</td>
<td>52.34(50.07/54.76)</td>
<td>54.90(55.40/54.30)</td>
</tr>
<tr>
<td><strong>Baseline</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>59.50(75.50/49.10)</td>
<td>21.77(46.91/14.18)</td>
<td>66.41(68.44/64.50)</td>
<td>43.34(58.05/34.59)</td>
<td>42.92(47.55/39.11)</td>
</tr>
<tr>
<td>AutoNER</td>
<td>67.00(75.21/60.40)</td>
<td><strong>26.10(43.26/18.69)</strong></td>
<td>67.18(64.63/69.95)</td>
<td><strong>51.39(48.82/54.23)</strong></td>
<td><strong>47.54(43.54/52.35)</strong></td>
</tr>
<tr>
<td>LRNT</td>
<td><strong>69.74(79.91/61.87)</strong></td>
<td>23.84(46.94/15.98)</td>
<td><strong>67.69(67.36/68.02)</strong></td>
<td>47.74(46.70/48.83)</td>
<td>46.21(45.60/46.84)</td>
</tr>
<tr>
<td><strong>Other Baseline</strong> (Reported Results)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KALM †</td>
<td>76.00(76.00/76.00)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ConNET °</td>
<td>75.57(84.11/68.61)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Our BOND Framework</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>46.61(53.11/41.52)</td>
<td>68.11(66.71/69.56)</td>
<td>59.11(60.14/58.11)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>BOND</td>
<td><strong>81.48(82.05/80.92)</strong></td>
<td>48.01(53.16/43.76)</td>
<td><strong>68.35(67.14/69.61)</strong></td>
<td><strong>65.74(67.37/64.19)</strong></td>
<td><strong>60.07(53.44/68.58)</strong></td>
</tr>
</tbody>
</table>
## Experiment: Main Result

### Table 2: Main Results on Testing Set: $F_1$ Score (Precision/Recall) (in %)

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Tweet</th>
<th>OntoNote5.0</th>
<th>Webpage</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity Types</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KB Matching</td>
<td>71.40(81.13/63.75)</td>
<td>35.83(40.34/32.22)</td>
<td>59.51(63.86/55.71)</td>
<td>52.45(62.59/45.14)</td>
<td>47.76(47.90/47.63)</td>
</tr>
<tr>
<td><strong>Fully-Supervised</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>90.11(89.14/91.10)</td>
<td>52.19(51.76/52.63)</td>
<td>86.20(84.59/87.88)</td>
<td>72.39(66.29/79.73)</td>
<td>86.43(85.33/87.56)</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>91.21(91.35/91.06)</td>
<td>52.18(60.01/46.16)</td>
<td>86.17(85.99/86.36)</td>
<td>52.34(50.07/54.76)</td>
<td>54.90(55.40/54.30)</td>
</tr>
<tr>
<td><strong>Baseline</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>59.50(75.50/49.10)</td>
<td>21.77(46.91/14.18)</td>
<td>66.41(68.44/64.50)</td>
<td>43.34(58.05/34.59)</td>
<td>42.92(47.55/39.11)</td>
</tr>
<tr>
<td>AutoNER</td>
<td>67.00(75.21/60.40)</td>
<td>26.10(43.26/18.69)</td>
<td>67.18(64.63/69.95)</td>
<td>51.39(48.82/54.23)</td>
<td><strong>47.54(43.54/52.35)</strong></td>
</tr>
<tr>
<td>LRNT</td>
<td><strong>69.74(79.91/61.87)</strong></td>
<td>23.84(46.94/15.98)</td>
<td><strong>67.69(67.36/68.02)</strong></td>
<td>47.74(46.70/48.83)</td>
<td>46.21(43.60/46.84)</td>
</tr>
<tr>
<td><strong>Other Baseline</strong> (Reported Results)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KALM †</td>
<td>76.00( – / – )</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ConNET ‡</td>
<td>75.57(84.11/68.61)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Our BOND Framework</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td><strong>75.61(83.76/68.90)</strong></td>
<td>46.61(53.11/41.52)</td>
<td>68.11(66.71/69.56)</td>
<td>59.11(60.14/58.11)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>BOND</td>
<td>81.48(82.05/80.92)</td>
<td>48.01(53.16/43.76)</td>
<td>68.35(67.14/69.61)</td>
<td>65.74(67.37/64.19)</td>
<td>60.07(53.44/68.58)</td>
</tr>
</tbody>
</table>
### Experiment: Main Result

#### Table 2: Main Results on Testing Set: $F_1$ Score (Precision/Recall) (in %)

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Tweet</th>
<th>OntoNote5.0</th>
<th>Webpage</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity Types</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KB Matching</td>
<td>71.40(81.13/63.75)</td>
<td>35.83(40.34/32.22)</td>
<td>59.51(63.86/55.71)</td>
<td>52.45(62.59/45.14)</td>
<td>47.76(47.90/47.63)</td>
</tr>
<tr>
<td><strong>Fully-Supervised</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>90.11(89.14/91.10)</td>
<td>52.19(51.76/52.63)</td>
<td>86.20(84.59/87.88)</td>
<td>72.39(66.29/79.73)</td>
<td>86.43(85.33/87.56)</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>91.21(91.35/91.06)</td>
<td>52.18(60.01/46.16)</td>
<td>86.17(85.99/86.36)</td>
<td>52.34(50.07/54.76)</td>
<td>54.90(55.40/54.30)</td>
</tr>
<tr>
<td><strong>Baseline</strong> (Our implementation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>59.50(75.50/49.10)</td>
<td>21.77(46.91/14.18)</td>
<td>66.41(68.44/64.50)</td>
<td>43.34(58.05/34.59)</td>
<td>42.92(47.55/39.11)</td>
</tr>
<tr>
<td>AutoNER</td>
<td>67.00(75.21/60.40)</td>
<td>26.10(43.26/18.69)</td>
<td>67.18(64.63/69.95)</td>
<td>51.39(48.82/54.23)</td>
<td>47.54(43.54/52.35)</td>
</tr>
<tr>
<td>LRNRT</td>
<td>69.74(79.91/61.87)</td>
<td>23.84(46.94/15.98)</td>
<td>67.69(67.36/68.02)</td>
<td>47.74(46.70/48.83)</td>
<td>46.21(45.60/46.84)</td>
</tr>
<tr>
<td><strong>Other Baseline</strong> (Reported Results)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KALM †</td>
<td>76.00(–/–)</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
</tr>
<tr>
<td>ConNET ‡</td>
<td>75.57(84.11/68.61)</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
</tr>
<tr>
<td><strong>Our BOND Framework</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>46.61(53.11/41.52)</td>
<td>68.11(66.71/69.56)</td>
<td>59.11(60.14/58.11)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>BOND</td>
<td>81.48(82.05/80.92)</td>
<td>48.01(53.16/43.76)</td>
<td>68.35(67.14/69.61)</td>
<td>65.74(67.37/64.19)</td>
<td>60.07(53.44/68.58)</td>
</tr>
</tbody>
</table>
### Experiment: Ablation

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Precision/Recall)</td>
<td>(Precision/Recall)</td>
</tr>
<tr>
<td><strong>Stage I</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>Stage I w/o pre-train</td>
<td>36.66(37.49/35.75)</td>
<td>18.31(18.14/18.50)</td>
</tr>
<tr>
<td>Stage I w/o early stop</td>
<td>72.11(81.65/64.57)</td>
<td>49.68(48.67/50.74)</td>
</tr>
<tr>
<td>Stage I w/ MT</td>
<td>76.30(82.92/70.67)</td>
<td>46.68(49.82/43.91)</td>
</tr>
<tr>
<td>Stage I w/ VAT</td>
<td>76.38(82.58/71.04)</td>
<td>47.54(50.02/45.30)</td>
</tr>
<tr>
<td><strong>Stage I + Stage II</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOND †</td>
<td>77.28(83.42/71.98)</td>
<td>56.90(54.32/59.74)</td>
</tr>
<tr>
<td>BOND w/ soft</td>
<td>80.18(81.56/78.84)</td>
<td>58.64(58.29/65.79)</td>
</tr>
<tr>
<td>BOND w/ soft+high conf</td>
<td>81.48(82.05/80.92)</td>
<td>60.07(53.44/68.58)</td>
</tr>
<tr>
<td>BOND w/ reinit</td>
<td>78.17(85.05/72.31)</td>
<td>58.55(55.31/62.19)</td>
</tr>
<tr>
<td>BOND w/ soft+reinit</td>
<td>76.92(83.39/71.38)</td>
<td>54.09(50.72/57.94)</td>
</tr>
<tr>
<td>BOND w/ MT</td>
<td>77.16(82.79/72.25)</td>
<td>57.93(55.66/60.39)</td>
</tr>
<tr>
<td>BOND w/ VAT</td>
<td>77.64(85.62/70.69)</td>
<td>57.39(55.05/59.41)</td>
</tr>
</tbody>
</table>
**Experiment: Ablation**

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage I</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>Stage I w/o pre-train</td>
<td>36.66(37.49/35.75)</td>
<td>18.31(18.14/18.50)</td>
</tr>
<tr>
<td>Stage I w/o early stop</td>
<td>72.11(81.65/64.57)</td>
<td>49.68(48.67/50.74)</td>
</tr>
<tr>
<td>Stage I w/ MT</td>
<td>76.30(82.92/70.67)</td>
<td>46.68(49.82/43.91)</td>
</tr>
<tr>
<td>Stage I w/ VAT</td>
<td>76.38(82.58/71.04)</td>
<td>47.54(50.02/45.30)</td>
</tr>
<tr>
<td><strong>Stage I + Stage II</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOND †</td>
<td>77.28(83.42/71.98)</td>
<td>56.90(54.32/59.74)</td>
</tr>
<tr>
<td>BOND w/ soft</td>
<td>80.18(81.56/78.84)</td>
<td>58.64(58.29/65.79)</td>
</tr>
<tr>
<td>BOND w/ soft+high conf</td>
<td>81.48(82.05/80.92)</td>
<td>60.07(53.44/68.58)</td>
</tr>
<tr>
<td>BOND w/ reinit</td>
<td>78.17(85.05/72.31)</td>
<td>58.55(55.31/62.19)</td>
</tr>
<tr>
<td>BOND w/ soft+reinit</td>
<td>76.92(83.39/71.38)</td>
<td>54.09(50.72/57.94)</td>
</tr>
<tr>
<td>BOND w/ MT</td>
<td>77.16(82.79/72.25)</td>
<td>57.93(55.66/60.39)</td>
</tr>
<tr>
<td>BOND w/ VAT</td>
<td>77.64(85.62/70.69)</td>
<td>57.39(55.05/59.41)</td>
</tr>
</tbody>
</table>
## Experiment: Ablation

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage I</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>Stage I w/o pre-train</td>
<td>36.66(37.49/35.75)</td>
<td>18.31(18.14/18.50)</td>
</tr>
<tr>
<td>Stage I w/o early stop</td>
<td>72.11(81.65/64.57)</td>
<td>49.68(48.67/50.74)</td>
</tr>
<tr>
<td>Stage I w/ MT</td>
<td>76.30(82.92/70.67)</td>
<td>46.68(49.82/43.91)</td>
</tr>
<tr>
<td>Stage I w/ VAT</td>
<td>76.38(82.58/71.04)</td>
<td>47.54(50.02/45.30)</td>
</tr>
<tr>
<td><strong>Stage I + Stage II</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOND †</td>
<td>77.28(83.42/71.98)</td>
<td>56.90(54.32/59.74)</td>
</tr>
<tr>
<td>BOND w/ soft</td>
<td>80.18(81.56/78.84)</td>
<td>58.64(58.29/65.79)</td>
</tr>
<tr>
<td>BOND w/ soft+high conf</td>
<td>81.48(82.05/80.92)</td>
<td>60.07(53.44/68.58)</td>
</tr>
<tr>
<td>BOND w/ reinit</td>
<td>78.17(85.05/72.31)</td>
<td>58.55(55.31/62.19)</td>
</tr>
<tr>
<td>BOND w/ soft+reinit</td>
<td>76.92(83.39/71.38)</td>
<td>54.09(50.72/57.94)</td>
</tr>
<tr>
<td>BOND w/ MT</td>
<td>77.16(82.79/72.25)</td>
<td>57.93(55.66/60.39)</td>
</tr>
<tr>
<td>BOND w/ VAT</td>
<td>77.64(85.62/70.69)</td>
<td>57.39(55.05/59.41)</td>
</tr>
</tbody>
</table>
**Experiment: Ablation**

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03</th>
<th>Wikigold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>75.61(83.76/68.90)</td>
<td>51.55(49.17/54.50)</td>
</tr>
<tr>
<td>Stage I w/o pre-train</td>
<td>36.66(37.49/35.75)</td>
<td>18.31(18.14/18.50)</td>
</tr>
<tr>
<td>Stage I w/o early stop</td>
<td>72.11(81.65/64.57)</td>
<td>49.68(48.67/50.74)</td>
</tr>
<tr>
<td>Stage I w/ MT</td>
<td>76.30(82.92/70.67)</td>
<td>46.68(49.82/43.91)</td>
</tr>
<tr>
<td>Stage I w/ VAT</td>
<td>76.38(82.58/71.04)</td>
<td>47.54(50.02/45.30)</td>
</tr>
<tr>
<td>Stage I + Stage II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOND †</td>
<td>77.28(83.42/71.98)</td>
<td>56.90(54.32/59.74)</td>
</tr>
<tr>
<td>BOND w/ soft</td>
<td>80.18(81.56/78.84)</td>
<td>58.64(58.29/65.79)</td>
</tr>
<tr>
<td>BOND w/ soft+high conf</td>
<td>81.48(82.05/80.92)</td>
<td>60.07(53.44/68.58)</td>
</tr>
<tr>
<td><strong>BOND w/ reinit</strong></td>
<td>78.17(85.05/72.31)</td>
<td>58.55(55.31/62.19)</td>
</tr>
<tr>
<td>BOND w/ soft+reinit</td>
<td>76.92(83.39/71.38)</td>
<td>54.09(50.72/57.94)</td>
</tr>
<tr>
<td>BOND w/ MT</td>
<td>77.16(82.79/72.25)</td>
<td>57.93(55.66/60.39)</td>
</tr>
<tr>
<td>BOND w/ VAT</td>
<td>77.64(85.62/70.69)</td>
<td>57.39(55.05/59.41)</td>
</tr>
</tbody>
</table>
Experiment: Ablation

Figure 6: Learning Curves of BOND, BOND (w/ reinit), BOND (w/ soft) and BOND (w/ soft + reinit)
Experiment: Parameter Study

(a) The Early Stopping Time of Stage I – $T_1$ (b) The Early Stopping Time in Stage II – $T_3$ (c) The Confidence Threshold of Stage II – $\epsilon$

Figure 7: Parameter Study using CoNLL03: $F_1$, Precision, Recall on Testing Set (in %)
Experiment: Error Analysis

Figure 8: Recall of Knowledge Base Matching and different stages of BOND. The horizontal axis denotes the true entity type. The segments in a bar denote the portions of the entities being classified into different entity types.
Thank You! To Find Out More?

- Git: https://github.com/cliang1453/BOND