Neural Reranking for Named Entity Recognition

Jie Yang, Yue Zhang, Fei Dong

Singapore University of Technology and Design

September, 2017; Varna.
• Introduction
  – Named Entity Recognition (NER) task
  – Classical models

• Neural Reranking for NER

• Experiments
  – Baselines
  – Results
  – Examples

• Conclusion
• Named Entity Recognition task
  – Find named entities for texts

• Examples:

  [Barack Obama] _PER_ was born in [hawaii] _LOC_.

  Rare [Hendrix] _PER_ song draft sells for almost $17,000.

  [Volkswagen AG] _ORG_ won 77,719 registrations.


  The bank is a division of [First Union Corp] _ORG_.
Introduction

• Models?
  – Classification
  – Sequence labeling

• Examples:

The chairman of [the Federal Reserve]_ORG is [Ben Bernanke]_PERSON
• Models?
  – Classification
  – Sequence labeling: HMM, CRF

• Examples:

The chairman of [the Federal Reserve]_{ORG} is [Ben Bernanke]_{PERSON}
• **Representations?**
  – **Discrete Features**
  – **Neural Features**

**CRF Layer:**

<table>
<thead>
<tr>
<th></th>
<th>$t_{i-1}$</th>
<th>$t_i$</th>
<th>$t_{i+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discrete Features:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_{i-1}$</td>
<td>chairman</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_i$</td>
<td>of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_{i+1}$</td>
<td>the</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\mathbf{c}$ 
$\mathbf{t}$
Introduction

• **Representations?**
  
  – Discrete Features: $c_i, c_i c_{i-1}, c_{i-1} c_i c_{i+1}, ...$
  
  – Neural Features

CRF Layer:

Discrete Features:

- $c_{i-1}$
- Chairman
- $c_i$
- Of
- $c_{i+1}$
- The

...
• **Representations?**
  
  – Discrete Features \( c_i, c_i c_{i-1}, c_{i-1} c_i c_{i+1}, \ldots \)
  
  – Neural Features

- **CRF Layer:**

- **BiLSTM Layer:**

- **Neural Representation:**

  \[ \ldots, \text{chairman}, \text{of}, \text{the}, \ldots \]
Introduction

• Representations?
  – Discrete Features
  – Neural Features

\[ c_i, c_i c_{i-1}, c_{i-1} c_i c_{i+1}, \ldots \]

word embedding + LSTM

character embedding

CRF Layer:
BiLSTM Layer:
Neural Representation:

\[ t_{i-1} \quad t_i \quad t_{i+1} \]

\[ c_{i-1} \quad c_i \quad c_{i+1} \]

chairman  of  the  \ldots
• Challenges?

  – Internal Relations of Entity (with context)
    • “Germany beat Argentina 1:0.”  LOC beat ?
    • “Lin Dan beat Lee Chong Wei 2:0.”  ? beat PER
• Challenges?
  – Internal Relations of Entity (with context)
    • “Germany beat Argentina 1:0.”
    • “Lin Dan beat Lee Chong Wei 2:0.”
  – Sparsity of Words in Entity (OOV problem)
    • “Barack Obama was born in Hawaii.”
    • “Bruno Mars was born in Hawaii.”
Challenges?

- Internal Relations of Entity (with context)
  - “Germany beat Argentina 1:0.” LOC beat ?
  - “Lin Dan beat Lee Chong Wei 2:0.” ? beat PER

- Sparsity of Words in Entity (OOV problem)
  - “Barack Obama was born in Hawaii.”
  - “Bruno Mars was born in Hawaii.”

- Global Features
  - “A beat B to win FIFA World Cup final.”
  - “A beat B to win Olympic Badminton final.”
• Challenges?
  – Internal Relations of Entity (with context)
    • “Germany beat Argentina 1:0.” LOC beat ?
    • “Lin Dan beat Lee Chong Wei 2:0.” ? beat PER
  – Sparsity of Words in Entity (OOV problem)
    • “Barack Obama was born in Hawaii.”
    • “Bruno Mars was born in Hawaii.”
  – Global Features
    • “A beat B to win FIFA World Cup final.”
    • “A beat B to win Olympic Badminton final.”

• Solution ?
• **Neural Reranking model for NER**
  – Reranking model
    • Generate N-best sequences through base model
    • Select the best sequence using new reranking model
Reranking

• **Neural Reranking model for NER**
  
  – Reranking model
    
    • Generate N-best sequences through base model
    
    • Select the best sequence using new reranking model
  
  – Neural Reranking
    
    • Represent candidate results using neural features
• Neural Reranking Example

\[ S: \quad \text{Barack Obama was born in hawaii .} \]
Reranking

- Neural Reranking Example

S: Barack Obama was born in hawaii.

N-best Candidates

L1: B-PER O O O O B-LOC O
L2: B-LOC I-LOC O O O O O O
L3: B-PER I-PER O O O O B-LOC O
Ln: B-PER I-PER O O O O B-PER O
Reranking

- Neural Reranking Example

\[ S: \text{Barack Obama was born in hawaii}. \]

N-best Candidates

\[
\begin{align*}
L_1: & \quad B\text{-PER} \quad O \quad O \quad O \quad O \quad B\text{-LOC} \quad O \\
L_2: & \quad B\text{-LOC} \quad I\text{-LOC} \quad O \quad O \quad O \quad O \quad O \\
L_3: & \quad B\text{-PER} \quad I\text{-PER} \quad O \quad O \quad O \quad B\text{-LOC} \quad O \\
\ldots & \\
L_n: & \quad B\text{-PER} \quad I\text{-PER} \quad O \quad O \quad O \quad B\text{-PER} \quad O
\end{align*}
\]

Base Model

Reranking model to select best sequence
• **Our Reranking Model**
  
  – Replace recognized entity with its type
  – Non-entity words keep the same
  – Neural representation for candidate sequences
  – Turn to sentence classification problem

\[
S: \text{Barack Obama was born in hawaii .}
\]

\[
L_1: B-\text{PER} \ O \ O \ O \ B-\text{LOC} \ O \quad \rightarrow \quad C_1: \ PER \ \text{Obama was born in LOC .}
\]

\[
L_2: B-\text{LOC} \ I-\text{LOC} \ O \ O \ O \ O \ O \ O \quad \rightarrow \quad C_2: \ LOC \ was \ born \ in \ hawaii .
\]

\[
L_3: B-\text{PER} \ I-\text{PER} \ O \ O \ O \ B-\text{LOC} \ O \quad \rightarrow \quad C_3: \ PER \ was \ born \ in \ LOC .
\]

\[
\ldots
\]

\[
L_n: B-\text{PER} \ I-\text{PER} \ O \ O \ O \ B-\text{PER} \ O \quad \rightarrow \quad C_n: \ PER \ was \ born \ in \ \ldots \ PER .
\]
• **Advantages**
  
  – Learning sentence patterns automatically
    
      • “LOC beat LOC” > “LOC beat PER”
      
      • “PER beat PER” > “ORG beat PER”
  
  – Word sparsity is eliminated
      
    • “Barack Obama was born in Hawaii.” PER was born ...
    
    • “Bruno Mars was born in Hawaii.” PER was born ...
  
  – Global features captured using rich neural features
    
      • Word + Character representation
      
      • LSTM + CNN representation
Experiments

• Baselines
  – Discrete features with CRF
Experiments

• Baselines

– Discrete features with CRF

<table>
<thead>
<tr>
<th>Description</th>
<th>Feature Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>word grams</td>
<td>$w_i, w_iw_{i+1}$</td>
</tr>
<tr>
<td>shape, capital</td>
<td>$Sh(w_i), Ca(w_i)$</td>
</tr>
<tr>
<td>capital + word</td>
<td>$Ca(w_i)w_i$</td>
</tr>
<tr>
<td>connect word</td>
<td>$Co(w_i)$</td>
</tr>
<tr>
<td>capital + connect</td>
<td>$Ca(w_i)Co(w_i)$</td>
</tr>
<tr>
<td>cluster grams</td>
<td>$Cl(w_i), Cl(w_iw_{i+1})$</td>
</tr>
<tr>
<td>prefix, suffix</td>
<td>$Pr(w_i), Su(w_i)$</td>
</tr>
<tr>
<td>POS grams</td>
<td>$P(w_i, w_1w_{i+1}, w_{i-1}w_1w_{i+1})$</td>
</tr>
<tr>
<td>POS + word</td>
<td>$P(w_0)w_0$</td>
</tr>
</tbody>
</table>

Discrete Features
• **Baselines**
  
  – Discrete features with CRF
  
  – Neural features with CRF

![Neural Features Diagram]

Experiments
Experiments

- **Baselines**
  - Discrete features with CRF
  - Neural features with CRF

Character CNN representation
Experiments

• **Baselines**
  
  – Discrete features with CRF
  – Neural features with CRF

BiLSTM-CRF with character CNN feature
Experiments

• **Neural Reranking Structure**
  - LSTM + CNN representation for candidate sentences
  - Training
    • 5-folder experiment for generating training data
    • Sentence level regression for sentence level accuracy
    • Loss function
      \[
      J(\Theta) = \frac{1}{|D|} \sum_{(C_i, y_i) \in D} (y_i - s(C_i))^2 + \frac{\lambda}{2} ||\Theta||^2_2
      \]
      \[s(C_i): \text{reranker score}\]
      \[y_i: \text{sentence level accuracy}\]
• **Neural Reranking Structure**
  
  – Decoding
    
    • Mixture decoding strategy

    \[
    \hat{y}_i = \arg \max_{C_i \in C(S)} (\alpha s(C_i) + (1 - \alpha)p(L_i))
    \]

    \(s(C_i)\): reranker score
    
    \(p(L_i)\): probability of baseline output
    
    \(\alpha \in [0, 1]\): interpolation weight, tuned in dev data
### Experiments

- **Neural Reranking Structure**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-best</td>
<td>10</td>
<td>peepholes</td>
<td>no</td>
</tr>
<tr>
<td>wordDim</td>
<td>50</td>
<td>charDim</td>
<td>50</td>
</tr>
<tr>
<td>LSTM hidden</td>
<td>100</td>
<td>dropout</td>
<td>0.2</td>
</tr>
<tr>
<td>charCNN filter</td>
<td>50</td>
<td>batch size</td>
<td>128</td>
</tr>
<tr>
<td>wordCNN filter</td>
<td>100</td>
<td>λ</td>
<td>0.001</td>
</tr>
<tr>
<td>charCNN length</td>
<td>3</td>
<td>Adam $\beta_1$</td>
<td>0.1</td>
</tr>
<tr>
<td>wordCNN length</td>
<td>3</td>
<td>Adam $\beta_2$</td>
<td>0.999</td>
</tr>
<tr>
<td>learning rate</td>
<td>0.001</td>
<td>Adam $\epsilon$</td>
<td>1e-8</td>
</tr>
</tbody>
</table>

Hyper-parameters
Experiments

• Results

– Oracle result for discrete CRF (88.15% -> 97.13%)

![Graph showing Oracle scores vs Candidate number]

Oracle results of discrete CRF model
Experiments

• **Results**

  – Full model works better in long sentences

  Sentence selection accuracy (SSA) with sentence length

Experiments
Experiments

• **Results**

  – Help for identifying fix type of entities

  • PER/LOC/ORG are improved, MISC becomes worse

![F1-value comparison by entity types](image)
## Experiments

### Final Results

- Discrete baseline: reranker improves by 1.1%

<table>
<thead>
<tr>
<th>Model (%)</th>
<th>F1</th>
<th>ΔF1</th>
<th>SSA</th>
<th>ΔSSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>88.15</td>
<td>0</td>
<td>83.31</td>
<td>0</td>
</tr>
<tr>
<td>LSTM</td>
<td>88.75</td>
<td>0.60</td>
<td>84.41</td>
<td>1.10</td>
</tr>
<tr>
<td>LSTM+CNN</td>
<td>88.79</td>
<td>0.64</td>
<td>84.63</td>
<td>1.32</td>
</tr>
<tr>
<td>LSTM+char</td>
<td>88.93</td>
<td>0.78</td>
<td>84.69</td>
<td>1.38</td>
</tr>
<tr>
<td>Full model</td>
<td><strong>89.25</strong></td>
<td><strong>1.10</strong></td>
<td><strong>85.12</strong></td>
<td><strong>1.82</strong></td>
</tr>
</tbody>
</table>

Final result for discrete baseline
Final Results

- Discrete baseline:

<table>
<thead>
<tr>
<th>Discrete Model (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kazama and Torisawa (2007)</td>
<td>88.02</td>
</tr>
<tr>
<td>Suzuki and Isozaki (2008)</td>
<td>89.92</td>
</tr>
<tr>
<td>Nguyen et al. (2010)</td>
<td>88.16</td>
</tr>
<tr>
<td>Ratinov and Roth (2009)</td>
<td>88.55</td>
</tr>
<tr>
<td>Ratinov and Roth (2009)*</td>
<td>90.57</td>
</tr>
<tr>
<td>Luo et al. (2015)</td>
<td>91.20</td>
</tr>
<tr>
<td>Discrete baseline</td>
<td>88.13</td>
</tr>
<tr>
<td>Our reranker</td>
<td>89.25</td>
</tr>
</tbody>
</table>

State-of-the-art discrete systems
Experiments

- **Final Results**
  
  - Neural baseline: reranker gives best results

<table>
<thead>
<tr>
<th>Neural Model (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collobert et al. (2011)</td>
<td>89.59</td>
</tr>
<tr>
<td>Passos et al. (2014)</td>
<td>90.90</td>
</tr>
<tr>
<td>Huang et al. (2015)</td>
<td>90.10</td>
</tr>
<tr>
<td>Chiu and Nichols (2016)</td>
<td>90.77</td>
</tr>
<tr>
<td>Lample et al. (2016)</td>
<td>90.94</td>
</tr>
<tr>
<td>Ma and Hovy (2016)</td>
<td>91.21</td>
</tr>
<tr>
<td>Neural baseline</td>
<td></td>
</tr>
<tr>
<td>Our reranker</td>
<td>91.25</td>
</tr>
</tbody>
</table>

State-of-the-art neural systems
Experiments

- **Examples**
  - Learns the right sentence pattern

<table>
<thead>
<tr>
<th>Baseline 1</th>
<th>Reranker 1</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Baseline 2</th>
<th>Reranker 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>West [Indian] \textit{MISC} all-rounder [Phil Simmons] \textit{PER} took four ...</td>
<td>[West Indian] \textit{MISC} all-rounder [Phil Simmons] \textit{PER} took four ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline 3</th>
<th>Reranker 3</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Baseline 4</th>
<th>Reranker 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>... prisoners are held in [Rangoon] \textit{LOC} ’s [Insein Prison] \textit{PER} .</td>
<td>... prisoners are held in [Rangoon] \textit{LOC} ’s [Insein Prison] \textit{LOC} .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline 5</th>
<th>Reranker 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>[PAKISTAN] \textit{LOC} WIN TOSS , PUT [ENGLAND] \textit{ORG} INTO BAT.</td>
<td>[PAKISTAN] \textit{LOC} WIN TOSS , PUT [ENGLAND] \textit{LOC} INTO BAT.</td>
</tr>
</tbody>
</table>
Conclusion

• Proposed a neural reranker for NER
• Improved most on fixed type entities
• Improved more on longer sentences
• Achieved state-of-the-art result on CoNLL03
Thanks!

Code available @ https://github.com/jiesutd/RerankNER