Improving Coreference Resolution by Learning Entity-Level Distributed Representations

K. Clark and C. Manning, ACL 2016
Coreference from clustering – Why?

• - Learns entity-level
  • Bill Clinton says...
  • Clinton..., she is very happy to be home.
  • {Bill Clinton}, {Clinton, she}. 
Model – Overall Design
Model – Mention Pair Encoder

• Obama says the U.S. government has killed Bin Laden.
  • Obama: {NA}
  • U.S. government: {Obama}
  • Bin Laden: {U.S. government, Obama}
Model – Mention Pair Encoder

Mention-Pair Representation $r_m$

Hidden Layer $h_2$
ReLU($W_3h_2 + b_3$)

Hidden Layer $h_1$
ReLU($W_2h_1 + b_2$)

Input Layer $h_0$
ReLU($W_1h_0 + b_1$)

Candidate Antecedent Embeddings
Candidate Antecedent Features
Mention Embeddings
Mention Features
Pair and Document Features
Model – Mention Pair Encoder

• Mention Features:
  • Type / position /...

• Pair&Document Features:
  • Genre / Distance / Speaker / String Match /

• Mention Embeddings:
  • head word / dependency parent / first(last word) / two preceding(following) words / averaged five preceding(following) words / averaged all words(mention,sentence,document) /
Model – Cluster Pair Encoder

Cluster-Pair Representation
\( r_c(c_1, c_2) \)

Mention-Pair Representations
\( R_m(c_1, c_2) \)

Pooling

\( c_1 \)
\( m_1^1 \)
\( m_1^2 \)

\( c_2 \)
\( m_2^1 \)
\( m_2^2 \)
Model – Mention Pair Ranker

\[ \hat{t}_i = \arg\max_{t \in \mathcal{T}(m_i)} s_m(t, m_i) \]

\[ \sum_{i=1}^{N} \max_{a \in \mathcal{A}(m_i)} \Delta(a, m_i) \left(1 + s_m(a, m_i) - s_m(\hat{t}_i, m_i)\right) \]

\[ \Delta(a, m_i) = \begin{cases} 
\alpha_{\text{FN}} & \text{if } a = \text{NA} \land \mathcal{T}(m_i) \neq \{\text{NA}\} \\
\alpha_{\text{FA}} & \text{if } a \neq \text{NA} \land \mathcal{T}(m_i) = \{\text{NA}\} \\
\alpha_{\text{WL}} & \text{if } a \neq \text{NA} \land a \notin \mathcal{T}(m_i) \\
0 & \text{if } a \in \mathcal{T}(m_i) 
\end{cases} \]
Model – Cluster Ranking

\[ \pi(\text{MERGE}[c_m, c]|x) \propto e^{s_c(c_m, c)} \]

\[ \pi(\text{PASS}|x) \propto e^{s_{NA}(m)} \]

• Easy First
  • Make easy decisions first
  • Delay hard ones to the last
  • Intuition?

• - Deep Learning to Search
  • Decisions made based on previous decisions
Model – Deep Learning to Search

for $i = 1$ to num_epochs do
    Initialize the current training set $\Gamma = \emptyset$
    for each example $(x, y) \in \mathcal{D}$ do
        Run the policy $\pi$ to completion from start state $x$ to obtain a trajectory of states \{${x_1, x_2, \ldots, x_n}$\}
        for each state $x_i$ in the trajectory do
            for each possible action $u \in U(x_i)$ do
                Execute $u$ on $x_i$ and then run the reference policy $\pi_{\text{ref}}$ until reaching an end state $e$
                Assign $u$ a cost by computing the loss on the end state: $l(u) = \mathcal{L}(e, y)$
            end for
        end for
        Add the state $x_i$ and associated costs $l$ to $\Gamma$
    end for
end for
Update $\pi$ with gradient descent, minimizing $\sum_{(x, l) \in \Gamma} \sum_{u \in U(x)} \pi(u|x)l(u)$. 
Model – Deep Learning to Search

- Run current policy from the start state to end
- Compute loss and update policy with gradient descent
- Expose to mistake, learns how to cope
# Results

<table>
<thead>
<tr>
<th>Model</th>
<th>English $F_1$</th>
<th>Chinese $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td>65.52</td>
<td>64.41</td>
</tr>
<tr>
<td>- MENTION</td>
<td>-1.27</td>
<td>-0.74</td>
</tr>
<tr>
<td>- GENRE</td>
<td>-0.25</td>
<td>-2.91</td>
</tr>
<tr>
<td>- DISTANCE</td>
<td>-2.42</td>
<td>-2.41</td>
</tr>
<tr>
<td>- SPEAKER</td>
<td>-1.26</td>
<td>-0.93</td>
</tr>
<tr>
<td>- MATCHING</td>
<td>-2.07</td>
<td>-3.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>English $F_1$</th>
<th>Chinese $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td>66.01</td>
<td>64.86</td>
</tr>
<tr>
<td>- PRETRAINING</td>
<td>-5.01</td>
<td>-6.85</td>
</tr>
<tr>
<td>- EASY-FIRST</td>
<td>-0.15</td>
<td>-0.12</td>
</tr>
<tr>
<td>- L2S</td>
<td>-0.32</td>
<td>-0.25</td>
</tr>
</tbody>
</table>

Table 1: CoNLL $F_1$ scores of the mention-ranking model on the dev sets without mention, document genre, distance, speaker, and string matching hand-engineered features.

Table 3: CoNLL $F_1$ scores of the cluster-ranking model on the dev sets with various ablations.
## Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CoNLL 2012 English Test Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark and Manning (2015)</td>
<td>76.12</td>
<td>69.38</td>
<td>72.59</td>
<td>65.64</td>
<td>56.01</td>
<td>60.44</td>
<td>59.44</td>
<td>52.98</td>
<td>56.02</td>
<td>63.02</td>
</tr>
<tr>
<td>Peng et al. (2015)</td>
<td>–</td>
<td>–</td>
<td>72.22</td>
<td>–</td>
<td>–</td>
<td>60.50</td>
<td>–</td>
<td>–</td>
<td>56.37</td>
<td>63.03</td>
</tr>
<tr>
<td>Wiseman et al. (2015)</td>
<td>76.23</td>
<td>69.31</td>
<td>72.60</td>
<td>66.07</td>
<td>55.83</td>
<td>60.52</td>
<td>59.41</td>
<td>54.88</td>
<td>57.05</td>
<td>63.39</td>
</tr>
<tr>
<td>Wiseman et al. (2016)</td>
<td>77.49</td>
<td>69.75</td>
<td>73.42</td>
<td>66.83</td>
<td>56.95</td>
<td>61.50</td>
<td>62.14</td>
<td>53.85</td>
<td>57.70</td>
<td>64.21</td>
</tr>
<tr>
<td>NN Mention Ranker</td>
<td>79.77</td>
<td>69.10</td>
<td>74.05</td>
<td>69.68</td>
<td>56.37</td>
<td>62.32</td>
<td>63.02</td>
<td>53.59</td>
<td>57.92</td>
<td>64.76</td>
</tr>
<tr>
<td>NN Cluster Ranker</td>
<td>78.93</td>
<td>69.75</td>
<td><strong>74.06</strong></td>
<td>70.08</td>
<td>56.98</td>
<td><strong>62.86</strong></td>
<td>62.48</td>
<td>55.82</td>
<td><strong>58.96</strong></td>
<td><strong>65.29</strong></td>
</tr>
</tbody>
</table>
Takeaway

• Clustering Coreference – Learns entity level information
• Deep learns policy with easy-first