Learning-based Multi-Sieve Co-reference Resolution with Knowledge.

Lev Ratinov, Dan Roth

Special thanks to: Nick Rizzolo, Kai-Wei Chang, EMNLP reviewers.
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
“After the vessel suffered a catastrophic torpedo detonation, Kursk sank in the waters of Barents Sea with all hands lost. Though rescue attempts were offered by a nearby Norwegian ship, Russia declined the initial offers, and all 118 sailors and officers aboard Kursk perished.”
“After the vessel suffered a catastrophic torpedo detonation, Kursk sank in the waters of Barents Sea with all hands lost. Though rescue attempts were offered by a nearby Norwegian ship, Russia declined the initial offers, and all 118 sailors and officers aboard Kursk perished.”
Clustering – recap.

\[ F_\theta(m_i, m_j) \rightarrow [0,1] \]
Desired Pairwise Prediction

[Diagram with nodes labeled [vessel], [Norwegian ship], [Kursk], and edges with weights 0.4, 0.51, 0.1, 0.1, 0.51, 0.99]
Training $\mathcal{F}_\theta(m_i, m_j) \rightarrow [0,1]$
Training $\mathcal{F}_\theta(m_1, m_j) \rightarrow [0,1]$
Training $F_\theta(m_i, m_j) \rightarrow [0,1]$
Training $\mathcal{F}_\theta(m_1, m_2) \rightarrow [0,1]$

$\mathcal{F}_\theta(m_1, m_2) \rightarrow 1$
$\mathcal{F}_\theta(m_1, m_3) \rightarrow 0$
$\mathcal{F}_\theta(m_2, m_3) \rightarrow 0$
$\mathcal{F}_\theta(m_2, m_4) \rightarrow 1$
$\mathcal{F}_\theta(m_3, m_4) \rightarrow 0$
Training $F_\theta(\neg m_i, m_j) \rightarrow [0,1]$
Inference – Best Link

[vessel]  $m_1$

[Norwegian ship]  $m_3$

[Kursk]  $m_2$

[Kursk]  $m_4$
Inference – Best Link

\[ F_\theta(m_1, m_2) = 0.51 \]
Inference – Best Link
Inference – Best Link

![](image)

1. **[vessel]**
   - $m_1$
   - 0.51

2. **[Kursk]**
   - $m_2$
   - 0.2

3. **[Norwegian ship]**
   - $m_3$

4. **[Kursk]**
   - $m_4$
Inference – Best Link

Keep only the “best co-ref link”
Inference – Best Link

[vessel] \( m_1 \) [Norwegian ship] \( m_3 \)

[Kursk] \( m_2 \) [Kursk] \( m_4 \)
Inference – Best Link

[vessel]

[Norwegian ship]

[Kursk]

[Kursk]
Inference – Best Link

[vessel] \( m_1 \) [Norwegian ship] \( m_3 \)

[Kursk] \( m_2 \) [Kursk] \( m_4 \)
Inference – Best Link

[vessel]  \( m_1 \)  [Norwegian ship]  \( m_3 \)

[Kursk]  \( m_2 \)  [Kursk]  \( m_4 \)
Inference – Best Link

Transitive closure
Why is co-ref a hard problem?
Features/Knowledge!

WordNet

No Knowledge

Missing relations:
• Nearby
• Rescue offers
Our focus: using Wikipedia

Step 1: Context-sensitive disambiguation

Step 2: Attribute extraction

- Nationality: Russian/Soviet
- Gender: Inanimate
- Fine-grained types: ship, submarine, incident
- Coarse-grained types: MAN_MADE_OBJECT
Our focus: using Wikipedia

Step 1: Context-sensitive disambiguation

Step 2: Attribute extraction
Co-ref Algorithm

Input: a set of mentions $m_1, \ldots, m_N$
Co-ref Algorithm

Input: a set of mentions $m_1$, ..., $m_N$

1. For each $m_i$,
   1.1. Assign $m_i$ a Wikipedia page.
   1.2. Assign $m_i$ a set of knowledge attributes.
Co-ref Algorithm

Input: a set of mentions $m_1, ..., m_N$

1. For each $m_i$,
   1.1. Assign $m_i$ a Wikipedia page.
   1.2. Assign $m_i$ a set of knowledge attributes.

2. Let Q be a queue of mention pairs $\{(m_i, m_j)\}$ in “easy-first” order.
   2.1. Let $G$ be a partial solution.
   2.2. As long as Q is not empty.
       2.2.1. Extract a pair $p=(m_i, m_j)$ from Q.
       2.2.2. Decide which classifier to use for the pair.
       2.2.3. Apply the classifier and update $G$.

Contributions of this work

• Using Wikipedia to assign a set of knowledge attributes to mentions in a context-sensitive way.
  – Context-sensitive disambiguation to Wikipedia
    http://cogcomp.cs.illinois.edu/page/software_view/Wikifier
  – Extracting attributes from Wikipedia pages.
  – Assigning attributes to mentions.
  – Using the attributes in compatibility metric.

• Porting the strengths of rule-based systems into a ML framework.
  – Learning-based sieves. Mention pairs can belong to more than one sieve.
  – Sieves can make contradicting predictions.
  – Later sieves leverage decisions of lower sieves.
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
Extracting knowledge from Wikipedia

Russian submarine K-141 Kursk

From Wikipedia, the free encyclopedia

K-141 Kursk was an Oscar-II class nuclear-powered cruise missile submarine of the Russian Navy, lost with all hands when it sank in the Barents Sea on August 12, 2000. Kursk, full name Атомная подводная лодка «Курск», which translated, means the nuclear powered submarine "Kursk" [АПЛ "Курск"] in Russian, was a Project 949A Антей (Antey, Antaeus but was also known by its NATO reporting name of Oscar II). It was named after the Russian city Kursk.

Categories: Oscar class submarines | Ships built in Russia | 1994 ships | Ships of the Russian Northern Fleet
K-141 Kursk accident | Submarine accidents caused by torpedoes | Sunken nuclear submarines
Shipwrecks in the Barents Sea | Lost submarines of Russia | Maritime incidents in 2000
Extracting knowledge from Wikipedia

Text

Russian submarine K-141 Kursk

From Wikipedia, the free encyclopedia

K-141 Kursk was an Oscar-II class nuclear-powered cruise missile submarine of the Russian Navy, lost with all hands when it sank in the Barents Sea on August 12, 2000. Kursk, full name Атомная подводная лодка «Курск», which translated, means the nuclear powered submarine "Kursk" [АПЛ "Курск"] in Russian, was a Project 949А Антей (Antey, Antaeus but was also known by its NATO reporting name of Oscar II). It was named after the Russian city Kursk,

Categories

Categories: Oscar class submarines | Ships built in Russia | 1994 ships | Ships of the Russian Northern Fleet | K-141 Kursk accident | Submarine accidents caused by torpedoes | Sunken nuclear submarines | Shipwrecks in the Barents Sea | Lost submarines of Russia | Maritime incidents in 2000
3 main types of knowledge

1) **Fine-grained semantic categories.**
2) **Gender.**
3) **Nationality.**

**Russian submarine K-141 Kursk**

From Wikipedia, the free encyclopedia

**K-141 Kursk** was an Oscar-II class nuclear-powered cruise missile submarine of the Russian Navy, lost with all hands when it sank in the Barents Sea on August 12, 2000. Kursk, full name Атомная подводная лодка «Курск», which translated, means the nuclear powered submarine "Kursk" [АПЛ "Курск"] in Russian, was a Project 949А Антей (Antey, Antaeus but was also known by its NATO reporting name of Oscar II). It was named after the Russian city Kursk.

---

Categories: Oscar class submarines | Ships built in Russia | 1994 ships | Ships of the Russian Northern Fleet
K-141 Kursk accident | Submarine accidents caused by torpedoes | Sunken nuclear submarines
Shipwrecks in the Barents Sea | Lost submarines of Russia | Maritime incidents in 2000
Fine-grained semantic categories.

Categories: Oscar class submarines | Ships built in Russia | 1994 ships | Ships of the Russian Northern Fleet
K-141 Kursk accident | Submarine accidents caused by torpedoes | Sunken nuclear submarines
Shipwrecks in the Barents Sea | Lost submarines of Russia | Maritime incidents in 2000

• Related work: (Suchanek et al., 2007; Nastase and Strube, 2008).

• We were looking for higher precision. Exploiting the structure:
  <nounphrase><relation-token><noun-phrase>
  – Remove the text succeeding a set of carefully chosen relation tokens.
  – Keep only the tokens labeled as NNS by POS tagger.
    “Institute for Advanced Study faculty” → “Institute” → “Institute”
  – Remove rare category tokens (assigned to less than 100 titles)

• 2088 fine-grained entity types.
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
Injecting knowledge into mentions

- Related work: *(Rahman and Ng, 2011; Ponzetto and Strube, 2006)*

- Context-sensitive disambiguation with GLOW:
  [http://cogcomp.cs.illinois.edu/page/software_view/Wikifier](http://cogcomp.cs.illinois.edu/page/software_view/Wikifier)
  - Named Entities with Wikipedia Disambiguation:

  - Mentions without GLOW mapping to Wikipedia:
    - Heuristic for removing “attributes” from a mention:
      - “{[Joseph Conrad Parkhurst], who founded the motorcycle magazine Cycle World in 1962 }, ...” → [Joseph Conrad Parkhurst]
    - If the head/extent contains a words from the 2088 fine-grained entity types extracted from Wikipedia, use the rightmost word (prefer head).
      - “{current Cycle World publisher [Larry Little]}”
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
Intuitions.

- (Easy/precision)-first inference + Stacking

- Projection - not all mention pairs have a similar decision boundary.
  - Splitting the data and using a different classifier for each split.
  - Vilalta & Rish. A decomposition of classes via clustering to explain and improve Naïve Bayes. ECML 2003.
Easy-first : Example

[Kramnik] has shocked the chess world by beating the [world champion], [Garry Kasparov].
Projection: Example

[Mubarak]_{m2}, wife of Egyptian President [Hosni Mubarak]_{m3}, and one of the conference organizers, said they must find ways to ...

String edit distance:
When two NERs are in the same sentence, small edit distance indicates non-co-reference.
Selected Sieves

<table>
<thead>
<tr>
<th>Sieve</th>
<th>Trained on All Data</th>
<th>Sieve-specific Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllSentencePairs</td>
<td>61.37</td>
<td>67.46</td>
</tr>
<tr>
<td>ClosestNonProDiffSent</td>
<td>60.71</td>
<td>63.33</td>
</tr>
<tr>
<td>NonProSameSentence</td>
<td>62.97</td>
<td>63.80</td>
</tr>
<tr>
<td>NerMentionsDiffSent</td>
<td>86.44</td>
<td>87.12</td>
</tr>
<tr>
<td>SameSentenceOneNer</td>
<td>64.10</td>
<td>68.88</td>
</tr>
<tr>
<td>Adjacent</td>
<td>71.00</td>
<td>78.80</td>
</tr>
<tr>
<td>SameSenBothNer</td>
<td>75.30</td>
<td>73.75</td>
</tr>
<tr>
<td>Nested</td>
<td>76.11</td>
<td>79.00</td>
</tr>
</tbody>
</table>

Table 1: F1 performance on co-referent mention pairs by sieve type when trained with all data versus sieve-specific data only.
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
Intermediate clustering (IC) features

- Let the set of sieves/classifiers be: \( \{C_1, C_2, \ldots, C_N\} \)
- Let \( R_i^+(m) \) be the set of mentions co-refed with \( m \) at sieve \( i \). Similarly, \( R_i^-(m) \) are non-corefed mentions.
  - \( E_i^+(m) = \bigcup_{j=1}^{i-1} R_j^+(m) \); \( E_i^-(m) = \bigcup_{j=1}^{i-1} R_j^-(m) \)

\[
IC_i^R(m_j, m_k) = \begin{cases} 
-1 & m_j \in R_{i-1}^-(m_k) \\
+1 & m_j \in R_{i-1}^+(m_k) \\
0 & \text{Otherwise}
\end{cases}
\]

\[
IC_i^E(m_j, m_k) = \begin{cases} 
-1 & m_j \in E_{i-1}^-(m_k) \\
+1 & m_j \in E_{i-1}^+(m_k) \\
0 & \text{Otherwise}
\end{cases}
\]
What we basically want to know

Stacking: after applying the sieves [1..i]:

a) Do mentions j and k fall in the same cluster through transitive closure?
What we basically want to know

Stacking: after applying the sieves [1..i]:

a) Do mentions j and k fall in the same cluster through transitive closure?

b) What is the minimal number of sieves we need to apply for (a) to happen?
What we basically want to know

Stacking: after applying the sieves [1..i]:

a) Do mentions j and k fall in the same cluster through transitive closure?

b) What is the minimal number of sieves we need to apply for (a) to happen?

c) Are there any negative direct non-coref predictions between j and k? At what sieve? What if we apply transitive closure on positive/negative predictions?
What we basically want to know

Stacking: after applying the sieves [1..i]:

a) Do mentions j and k fall in the same cluster through transitive closure?

b) What is the minimal number of sieves we need to apply for (a) to happen?

c) Are there any negative direct non-coref predictions between j and k? At what sieve? What if we apply transitive closure on positive/negative predictions?

d) What about surface forms? E.g. were there any co-ref predictions between “Kursk” and “vessel” elsewhere in the document? At what sieve?
Outline

• Introduction.
  – Task definition.
  – Training and inference.
  – System overview.

• Knowledge.
  – Extracting knowledge from Wikipedia.
  – Context-sensitive injection of knowledge to mentions.

• Learning-based Multi-Sieve Co-reference Resolution.
  – Why/how.
  – Sieve-based “entity features”.

• Results.
Methodology

• Off-the shelf system of Bengston.
• ACE 2004 dataset, stripped of transcripts.
• Two performance metrics
  – F1 on co-referent mention pairs (sieve by sieve)
Utility of knowledge and prediction features (F1 on co-referent mention pairs) by inference sieves.

<table>
<thead>
<tr>
<th></th>
<th>(B) Baseline</th>
<th>(B) + Knowledge</th>
<th>(B) + Predictions</th>
<th>(B) + Knowledge + Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopSieve</td>
<td>66.58</td>
<td>69.08</td>
<td>68.77</td>
<td>70.43</td>
</tr>
<tr>
<td>AllSentencePairs</td>
<td>67.46</td>
<td>71.79</td>
<td>69.59</td>
<td>73.50</td>
</tr>
<tr>
<td>ClosestNonProDiffSent</td>
<td>63.33</td>
<td>65.62</td>
<td>65.57</td>
<td>70.76</td>
</tr>
<tr>
<td>NonProSameSentence</td>
<td>63.80</td>
<td>69.62</td>
<td>67.03</td>
<td>71.11</td>
</tr>
<tr>
<td>NerMentionsDiffSent</td>
<td>87.12</td>
<td>88.23</td>
<td>88.68</td>
<td>89.07</td>
</tr>
<tr>
<td>SameSentenceOneNer</td>
<td>68.88</td>
<td>70.58</td>
<td>67.89</td>
<td>73.17</td>
</tr>
<tr>
<td>Adjacent</td>
<td>78.80</td>
<td>81.32</td>
<td>80.00</td>
<td>81.79</td>
</tr>
<tr>
<td>SameSenBothNer</td>
<td>73.75</td>
<td>80.50</td>
<td>77.21</td>
<td>80.98</td>
</tr>
<tr>
<td>Nested</td>
<td>79.00</td>
<td>83.59</td>
<td>80.65</td>
<td>83.37</td>
</tr>
</tbody>
</table>
Inference – Best Link

\[ F_\theta(m_i, m_j) > \alpha \]

[vessel] → [Norwegian ship]

[Kursk] → [Kursk]
End performance with Best Link.
Thank you.