Cross-Lingual Cross-Document Coreference with Entity Linking

Sean Monahan, John Lehmann,
Timothy Nyberg, Jesse Plymale, Arnold Jung
2010 Entity Linking Task

- Link entity mentions in text to Knowledge Base (KB)
  - Each entity mention is given a KB identifier
  - Non-clustering linker

The first **NATO** Secretary General, **Lord Ismay**, famously stated the organization's goal was "to keep the Russians out, the Americans in, and the Germans down". **John Smith** did not like this. He told this to his wife, **Jane Smith**.

The Berlin Plus agreement is a comprehensive package of agreements made between **NATO** and the **European Union** on 16 December 2002. **John Smith** disagreed with the first agreement.
2011 Entity Linking with NIL Clustering Task

- Additionally, cluster all of the remaining NILs
  - Perhaps the most important entities might be the ones you haven’t heard of yet
- Deductive approach: First link, then cluster remaining NILs

The first NATO Secretary General, Lord Ismay, famously stated the organization’s goal was "to keep the Russians out, the Americans in, and the Germans down". John Smith did not like this. He told this to his wife, Jane Smith.

The Berlin Plus agreement is a comprehensive package of agreements made between NATO and the European Union on 16 December 2002. John Smith disagreed with the first agreement.
2011 Entity Linking with NIL Clustering Task

- Alternate view: Cross-Document Coreference (CDC) approach
  - Cluster all mentions in text
  - Assign clusters a KB identifier
  - Inductive approach

The first NATO Secretary General, Lord Ismay, famously stated the organization's goal was "to keep the Russians out, the Americans in, and the Germans down". John Smith did not like this. He told this to his wife, Jane Smith.

The Berlin Plus agreement is a comprehensive package of agreements made between NATO and the European Union on 16 December 2002. John Smith disagreed with the first agreement.
Talk Overview

1. English Entity Linking (with NIL Clustering)
   - Made extensive use of 2010 Entity Linking System
     • Details in (Lehmann et al., 2010)
   - Focus on extending task to NIL clustering
     • 4-stage clustering algorithm
     • Show that our method:
       - Successfully performs NIL clustering
       - Improves linking accuracy on non-NIL entities
   - Improvements to 2010 entity linking algorithm (non-clustering)
2. Cross-Lingual Entity Linking with NIL Clustering
   – Two Approaches
     • Native Language Entity Linking
     • Translation with English Linking
2011 Entity Linking with NIL Clustering Components

• Necessary components

1. Synonymy
   • Determine entities likely to match
   • “National Security Council” → “NSC”

2. Polysemy
   • Extract features and cluster similar entities
   • “NSC” (Iran) ≠ “NSC” (Malaysia)

3. KB Linking / NIL Detection
   • Decide between the best KB identifier and NIL for each cluster
Approach

0. Preprocess each document
   - Includes entity links using the non-clustering linker
1. Group by similar names
2. Resolve polysemy with agglomerative clustering
3. Resolve synonymy by merging clusters
4. Link each cluster to the knowledge base
National Security Council

"We and other countries have expressed our concern to the Chinese," said a spokesman for the National Security Council, Gordon Johndroe.

Iran's National Security Council has announced that it will "suspend" the releasing of 15 British sailors and marines detained by Iranian forces on March 23.

The document "reflects the broad interagency effort under way in Iraq" according to an NSC spokesman Frederick Jones.
CDC: Stage 2

Cluster within the groups to resolve polysemy

“We and other countries have expressed our concern to the Chinese,” said a spokesman for the National Security Council, Gordon Johndroe.

Iran’s National Security Council has announced that it will "suspend" the releasing of 15 British sailors and marines detained by Iranian forces on March 23.

The document “reflects the broad interagency effort under way in Iraq” according to an NSC spokesman Frederick Jones.
CDC: Stage 2 Clustering Algorithm

Supervised hierarchical agglomerative clustering

- (Gooi and Allan, 1998)
- Balanced Data Set (Akbani et al., 2004)

\[ d(M_1, M_2) = \frac{1}{|M_1| \cdot |M_2|} \sum_{m_1 \in M_1} \sum_{m_2 \in M_2} d(m_1, m_2) \]

merge if \( d < \tau \)
CDC: Stage 2 Features

- Calculate similarity between mentions with a logistic regression classifier
  - (Mayfield et al., 2009)

### Key Features

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Type</td>
<td>Person, organization, etc…</td>
</tr>
<tr>
<td>Entity Links</td>
<td>Existence and confidence of same KB identifier (non-clustering)</td>
</tr>
<tr>
<td>Term Similarity</td>
<td>TFIDF weighted bag of words (Bagga/Baldwin 1998)</td>
</tr>
<tr>
<td>Local Context</td>
<td>E.g.: Actor Will Smith or Vice-President Will Smith</td>
</tr>
</tbody>
</table>
"We and other countries have expressed our concern to the Chinese," said a spokesman for the National Security Council, Gordon Johndroe.

Iran's National Security Council has announced that it will "suspend" the releasing of 15 British sailors and marines detained by Iranian forces on March 23.

The document "reflects the broad interagency effort under way in Iraq" according to an NSC spokesman Frederick Jones.
CDC: Stage 3 Model

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1 = 1$</td>
<td>If $m_1$ and $m_2$ have same KB identifier w/ confidence $&gt; \mu$</td>
</tr>
<tr>
<td>$I_2 = 1$</td>
<td>If $m_1$ and $m_2$ are embedded in a longer common phrase</td>
</tr>
</tbody>
</table>

\[
\sum_{m_1 \in M_1} \sum_{m_2 \in M_2} \alpha_k I_k(m_1, m_2) > \lambda, k \in (1, 2, \ldots)
\]
Stage 4: KB Identifier Generation

- Map each cluster to the knowledge base.
- Voting algorithm
  - Each entity link has a weight of 1

- Experimented with weighted links
English Entity Linking Submission

- 3 submissions
  - LCC3: Entity Linking with NIL Clustering System, without web access
    - Primary Evaluation
  - LCC1: Same as LCC3, with web access
  - LCC2: Changed model parameters to target precision

<table>
<thead>
<tr>
<th>Submission</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC3*</td>
<td>84.4</td>
<td>84.7</td>
<td>84.6</td>
</tr>
<tr>
<td>LCC1</td>
<td>86.7</td>
<td>87.1</td>
<td>86.9</td>
</tr>
<tr>
<td>LCC2</td>
<td>86.7</td>
<td>86.2</td>
<td>86.4</td>
</tr>
</tbody>
</table>

2011 KBP Submissions

- Attempting to improve precision ended up hurting recall
Inductive vs. Deductive Experiments

- **Inductive System**
  - Non-Clustering Linking as a feature
- **Deductive System**
  - Non-Clustering Linking as ground truth

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>MicroAvg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inductive</td>
<td>84.4</td>
<td>84.7</td>
<td>84.6</td>
<td>86.1</td>
</tr>
<tr>
<td>Deductive</td>
<td>84.2</td>
<td>83.7</td>
<td>84.0</td>
<td>85.7</td>
</tr>
</tbody>
</table>

2011 Eval Set

- +0.6 F
- +0.4 MicroAvg
Use of Non-Clustering Entity Linking Features

- Inductive system
  - Entity Links as a feature in Stages 2 and 3
  - Entity Links used to assign KB in Stage 4
- Without links as cluster features
  - Only uses entity links in Stage 4

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>MicroAvg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inductive</td>
<td>84.4</td>
<td>84.7</td>
<td>84.6</td>
<td>86.1</td>
</tr>
<tr>
<td>without links</td>
<td>82.1</td>
<td>83.2</td>
<td>82.7</td>
<td>84.7</td>
</tr>
</tbody>
</table>

2011 Eval Set

- +1.9 F
- +1.4 MicroAvg
2011 Non-Clustering Entity Linking Improvements

- Utilize Local Context
  - “Jim moved from Missouri to \textit{Springfield}, Illinois.”
  - “Joe lives in Atlanta, \textit{Georgia}”
- String normalization (diacritics)
  - “Jose” → “José”
- More precise candidate generation

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>MicroAvg</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>81.7</td>
<td>82.2</td>
<td>82.0</td>
<td>83.7</td>
</tr>
<tr>
<td>2011</td>
<td>84.4</td>
<td>84.7</td>
<td>84.6</td>
<td>86.1</td>
</tr>
</tbody>
</table>

2011 Eval Set

- +2.6 F
- +2.4 MicroAvg
Talk Overview

1. Entity Linking with NIL Clustering

2. Cross-lingual Entity Linking with NIL Clustering

   – Why is this task important?
   – Added Challenges
     • Linking Chinese entities
     • Clustering Chinese entities
     • Clustering English and Chinese entities
Cross-Language Linking Approaches

Chinese Wikipedia

Cross-Language Links

English Wikipedia

Definition

TAC Knowledge Base

Translation/Transliteration

Translation

Chinese Documents

Chinese Entity Linker

NKB

English Entity Linker

Cross-Language Linking Approaches
Native Language Knowledge Base Approach

• Link to the *Native Language Knowledge Base (NKB)*
• Wikipedia provides a useful knowledge base in many languages
  – 39 languages with > 100k pages
• Adapting our system to go from English to Chinese
  – See (Lehmann et al., 2010)
  – Candidate Generation
    • Wikipedia-based sources apply equally
    • Sources like acronym do not work
    • Search engine: “site:zh.wikipedia.org”
  – Candidate Ranking
    • Using low ambiguity link similarity
  – NIL Detection
    • Trained model for Chinese
  – Cluster Similarity
    • Context similarity using document context is language independent
    • Trained model for Chinese
Translation Approach

- Compared to NKB
  - Advantages: Can use our English linking system
  - Disadvantage: Translation fidelity
  - Unknown: Chinese vs. English entities
- Translate the query documents and queries (using Bing Translation API)
  - Use English system directly
- NKB performs 1.9 F better
- Combination algorithm
  - Run both systems, select most confident link, prefer non-NIL over NIL

<table>
<thead>
<tr>
<th>System</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>NKB</td>
<td>80.9</td>
</tr>
<tr>
<td>Translation</td>
<td>79.0</td>
</tr>
<tr>
<td>Voting</td>
<td>82.6</td>
</tr>
</tbody>
</table>

Score on Development Set

- +1.7 F
Cross-Lingual Scores

• 3 submissions
  – LCC1: NKB (no web)
    * Primary Evaluation
  – LCC2: NKB (with web)
  – LCC3: NKB (with web) combined Translation

<table>
<thead>
<tr>
<th>Submission</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Gain (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC1*</td>
<td>78.6</td>
<td>79.0</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>LCC2</td>
<td>80.7</td>
<td>81.2</td>
<td>80.9</td>
<td>+2.1</td>
</tr>
<tr>
<td>LCC3</td>
<td>78.8</td>
<td>81.3</td>
<td>80.0</td>
<td>+1.2</td>
</tr>
</tbody>
</table>

2011 KBP Cross-Lingual submissions

• +2.1 F with Web Features
• +1.2 F with Combined
Chinese vs. English linking

- Cross-lingual data contains both English and Chinese queries

<table>
<thead>
<tr>
<th>Submission</th>
<th>Combined</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC1 (no web)</td>
<td>82.4</td>
<td>84.6</td>
<td>81.3</td>
</tr>
<tr>
<td>LCC2</td>
<td>84.3</td>
<td>87.3</td>
<td>82.9</td>
</tr>
<tr>
<td>LCC3</td>
<td>83.9</td>
<td>87.5</td>
<td>82.2</td>
</tr>
</tbody>
</table>

Entity Linking Scores by language

- English several % better
- +1.6 F with Chinese Web
Development vs. Evaluation

- In development set, the combination system performed better than NKB system

<table>
<thead>
<tr>
<th>System</th>
<th>Dev Set</th>
<th>Eval Set</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>NKB</td>
<td>80.9</td>
<td>82.9</td>
<td>+2.0</td>
</tr>
<tr>
<td>Translation</td>
<td>79.0</td>
<td>79.8</td>
<td>+0.8</td>
</tr>
<tr>
<td>Voting</td>
<td>82.6</td>
<td>82.2</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

Entity Linking Scores (dev vs. eval)

- Both NKB and Translation performed better on evaluation set
Conclusions

• Inductive outperforms Deductive
• NKB outperforms Translation
  – Combined approach promising
• Clustering and Linking require little language customization
  – Could be an area for improvements
• Currently addressing scalability
  – Built a distributed clustering algorithm
    • Stores result in NoSQL database
    • Web front end
  – Working to scale to millions to documents
    • (Singh et al., 2011)
• Thank You!