Entity Clustering Across Languages

NAACL 2012 — Montreal

Spence Green* Nicholas Andrews# Matthew R. Gormley#
Mark Dredze# Christopher D. Manning*

*Stanford University
#Johns Hopkins University
How Many Different Ways Can You Spell ‘Gaddafi’?

Libyan leader Moammar Gaddafi has been in the headlines a lot this week — he's visiting the U.S. for the first time since he took power 40 years ago, and his arrival is not without controversy.

But politics aside, we've come across a list of the many different English spellings of Gaddafi's name. Because of the difficulty translating Arabic to English, there are several different translations — the Library of Congress lists 72 alternate spellings, and the New York Times, Associated Press and Xinhua news sources used 40 additional spellings between 1998 and 2008.

We've posted all 112 of them below...
One Entity, Many Names

Qaddafi, Muammar
Al-Gathafi, Muammar
al-Qadhafi, Muammar
Al Qathafi, Mu’ammar
Al Qathafi, Muammar
El Gaddafi, Moamar
El Kadhafi, Moammar
El Kazzafi, Moamer
El Qathafi, Mu’Ammar
One Entity, Many Names

Qaddafi, Muammar
Al-Gathafi, Muammar
al-Qadhafi, Muammar
Al Qathafi, Mu’ammar
Al Qathafi, Muammar
El Gaddafi, Moamar
El Kadhafi, Moammar
El Kazzafi, Moamer
El Qathafi, Mu’Ammer
One Entity, Many Names

Qaddafi, Muammar
Al-Gathafi, Muammar
al-Qadhafi, Muammar
Al Qathafi, Mu’ammar
Al Qathafi, Muammar
El Gaddafi, Moamar
El Kadhafi, Moammar
El Kazzafi, Moamer
El Qathafi, Mu’Ammar
Basic Task: Entity Clustering

Cluster co-referent entity mentions across a corpus
(documents and languages)
Basic Task: Entity Clustering

Cluster co-referent entity mentions across a corpus (documents and languages)

Clustering/disambiguation relies on:

- Mention similarity
- Context similarity
Entity Disambiguation: Mention Similarity

Qaddafi, Muammar
Al-Gathafi, Muammar
al-Qadhafi, Muammar
Al Qathafi, Mu’ammar
Al Qathafi, Muammar
El Gaddafi, Moamar
El Kadhafi, Moammar
El Kazzafi, Moamer
El Qathafi, Mu’Ammar
Entity Disambiguation: Context Similarity

Apple
Entity Disambiguation: Context Similarity

Apple

Apple Inc.

town in Lebanon

camel
Entity Disambiguation: Context Similarity

The **Apple** chief executive was former Beatles road manager Neil Aspinall...

Sentential context is usually required
Old: Entity Clustering Tasks

Within-doc coref

Peter said to himself ...
Old: Entity Clustering Tasks

**Within-doc coref**

Peter said to **himself** ...

**Cross-doc coref**

**doc1:** Peter Jones said ...

**doc2:** I told Mr. Jones ...

[Bagga and Baldwin, 1998]
[Baron and Freedman, 2008]

Peter said to himself ...
Old: Entity Clustering Tasks

Within-doc coref

Peter said to himself ...

Cross-doc coref

doc1: Peter Jones said ...

doc2: I told Mr. Jones ...

Entity Linking

Peter Jones
Mr. Jones

[Bagga and Baldwin, 1998]
[Baron and Freedman, 2008]

[McNamee et al., 2011]
[Rao et al., 2011]
New: Entity Clustering Across Languages

**Within-doc coref**

Peter said to himself ...

**Cross-doc coref**

**This paper**

[Bagga and Baldwin, 1998]
[Baron and Freedman, 2008]

[McNamee et al., 2011]
[Rao et al., 2011]
New: Entity Clustering Across Languages

The Apple chief executive was former Beatles road manager Neil Aspinall...

No knowledge base of entities
Why?

Crisis management

Arab Spring (2011)
- French, Arabic dialects
- Facebook, Twitter, blog...

Haiti earthquake (2010)
- Kreyol, English, French
- SMS, Twitter, blog...
Why?

Crisis management

Arab Spring (2011)
- French, Arabic dialects
- Facebook, Twitter, blog...

Haiti earthquake (2010)
- Kreyol, English, French
- SMS, Twitter, blog...

Other applications: machine translation, name search
Plan: Extend Existing Monolingual System

<table>
<thead>
<tr>
<th>Language</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>91.4</td>
</tr>
<tr>
<td>Arabic</td>
<td>89.8</td>
</tr>
<tr>
<td>English+Arabic</td>
<td>78.8</td>
</tr>
</tbody>
</table>
Cross-lingual Mention Similarity
Cross-lingual Context Similarity
Clustering Algorithms
Evaluation
Cross-lingual Mention Similarity

Cross-lingual Context Similarity

Clustering Algorithms

Evaluation
Within-language: Edit Distance

\[ m_i = \text{Muammar Qaddafi} \]
\[ m_j = \text{Moamer El Kazzafi} \]
\[ \text{sim}(m_i, m_j) = ? \]
Within-language: Edit Distance

\[ m_i = \text{Muammar Qaddafi} \]
\[ m_j = \text{Moamer El Kazzafi} \]
\[ \text{sim}(m_i, m_j) = \]?

**Algorithm:** Sorted Jaro-Winkler \[ [\text{Christen, 2006}] \]

1. Sort tokens
2. Compute Jaro-Winkler distance in \( O(|m_i| + |m_j|) \)
3. Evaluate \( \text{sim}(m_i, m_j) < \beta \)
Cross-language: Binary classifier

\[ m_i = \text{map(Apple)} = \text{abbl} \]

\[ m_j = \text{map(أبل)} = \text{abl} \]

\[ \text{sim}(m_i, m_j) = ? \]
Cross-language: Binary classifier

\[ m_i = \text{map}(\text{Apple}) = \text{abbl} \]

\[ m_j = \text{map}(\text{ابل}) = \text{abl} \]

\[ \text{sim}(m_i, m_j) = ? \]

**Algorithm:** Phonetic mapping + classification

1. Apply deterministic mapping \( \text{map}(\cdot) \)
2. Extract character-level features
3. Classify (Maxent)
Cross-language: Binary classifier

Training: parallel name list

- 97.1% accuracy on a held-out set

Phonetic mapping: think of Soundex

Best features: character bigrams
Cross-lingual Mention Similarity

Cross-lingual Context Similarity

Clustering Algorithms

Evaluation
## Context Similarity: Mapping techniques

<table>
<thead>
<tr>
<th>Method</th>
<th>Resource-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>High</td>
</tr>
<tr>
<td>Lexicon</td>
<td>Medium</td>
</tr>
<tr>
<td>Polylingual Topic Model</td>
<td>Low</td>
</tr>
</tbody>
</table>
# Context Similarity: Mapping techniques

<table>
<thead>
<tr>
<th>Method</th>
<th>Resource-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>High</td>
</tr>
<tr>
<td>Lexicon</td>
<td>Medium</td>
</tr>
<tr>
<td>Polylingual Topic Model</td>
<td>Low</td>
</tr>
</tbody>
</table>

- MT: Phrasal with NIST-09 data [Galley et al., 2009]
- Lexicon: 31k entries (web and LDC sources)
Polylingual Topic Model (PLTM) [Mimno et al., 2009]

- Words linked through cross-lingual priors
- Training: Wikipedia document tuples

Map context words to 1-best topics
Context Similarity: Polylingual Topic Model

[Mimno et al., 2009]
The Apple chief executive was former Beatles road manager Neil Aspinall...
Context Similarity

Bag of words / smoothed unigram distributions

Measure: Jensen-Shannon divergence
Cross-lingual Mention Similarity
Cross-lingual Context Similarity
Clustering Algorithms
Evaluation
Constraint-Based Clustering

Two algorithms:

1. Hierarchical clustering
2. Dirichlet process mixture model

Setup:

- Mention similarity as a hard constraint
- Cluster distance: context similarity
Constraint-Based Clustering

Muammar Qaddafi
El Kazzafi
context

al-Qadhafi
context

Apple
Apple Corps.
context
Constraint-Based Clustering

Muammar Qaddafi
El Kazzafi
context

al-Qadhafi
context

Apple
Apple Corps.
context

0.40
0.31
0.15
Constraint-Based Clustering

Muammar Qaddafi
El Kazzafi
context

al-Qadhafi
context

Apple
Apple Corps.

0.31
Cross-lingual Mention Similarity
Cross-lingual Context Similarity
Clustering Algorithms

Evaluation
Evaluation Corpus

Automatic Content Extraction (ACE) 2008 Arabic-English

We annotated **216 cross-lingual entities**

Genres:
1. broadcast conversation
2. broadcast news
3. meeting
4. newswire
5. telephone
6. usenet
7. weblog
ACE2008 Evaluation Corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>Docs</th>
<th>Tokens</th>
<th>Entities</th>
<th>Chains</th>
<th>Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>412</td>
<td>178k</td>
<td>2.6k</td>
<td>4.2k</td>
<td>9.2k</td>
</tr>
<tr>
<td>English</td>
<td>414</td>
<td>246k</td>
<td>2.3k</td>
<td>4.0k</td>
<td>9.1k</td>
</tr>
</tbody>
</table>

- **Chain** – set of mentions (within-doc)
- **Entity** – set of chains
Our models cluster **chains**

Evaluation:
- Gold chains
- Predicted chains from SERIF [Ramshaw et al., 2011]
Evaluation: Gold within-document processing

<table>
<thead>
<tr>
<th>B3</th>
<th>MT</th>
<th>Lexicon</th>
<th>PLTM</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3 (cross-lingual only)</td>
<td>78.8</td>
<td>58.4</td>
<td>54.5</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation: Gold within-document processing

<table>
<thead>
<tr>
<th>Method</th>
<th>B3</th>
<th>B3 (cross-lingual only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>85.4</td>
<td>78.8</td>
</tr>
<tr>
<td>Lexicon</td>
<td>80.4</td>
<td>66.4</td>
</tr>
<tr>
<td>PLTM</td>
<td>77.3</td>
<td>58.4</td>
</tr>
<tr>
<td>Baseline</td>
<td>70.1</td>
<td>54.5</td>
</tr>
</tbody>
</table>

In paper: CEAF, NVI
Automatic within-document processing

Bar chart showing performance metrics for MT, Lexicon, PLTM, and Baseline across different B3 values.

- MT: 76.7
- Lexicon: 76.0
- PLTM: 75.3
- Baseline: 67.0
Lexicon Clustering

Correct

tony blair
tyny blyr

khaled mashaal
khld mshal

Incorrect

NSA
CIA (En and Ar mentions)

republican party
≠
الحزب الجمهوري hzb jwmhwry
MT Clustering

Correct

NSA ≠ CIA

labour party = حزب العمل hzb aml

Incorrect

hamed bin khalifa al-thani ≠ حمد بن خليفة آل ثاني hmd bn khlfa thny
Conclusion

Within-doc coref

Peter said to himself ...

Cross-doc coref

doc1: Peter Jones said ...

doc2: I told Mr. Jones ...

[Rao et al., 2011]

Entity Linking

Mr. Jones

This paper

doc1: Peter Jones said ...

[Rao et al., 2011]

[Rao et al., 2011]

Peter Jones

Peter

[Raon and Freedman, 2008]

[McNamee et al., 2011]
Code and corpus:  spencegreen.com

thanks.
Future Work

Pairwise models don’t scale
  ▶ See [Rao et al., 2010] and [Singh et al., 2011]

Model at mention level
  ▶ See Nick Andrews’ talk at EMNLP!

Unified similarity measure
  ▶ Logistic regression did not generalize