Exposing Coptic entities
Automation, search and visualization

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Caroline T. Schroeder
University of the Pacific
cschroeder@pacific.edu

Lance Martin
Catholic University of America
71martin@cua.edu
copticscriptorium.org

Open source, open access research platform for Coptic language and literature

Tools, texts, collaborative environment

Interdisciplinary

Today: focus on entities
Thanks!

- This is joint work with project members and students:
  - Sichang Tu
  - Elizabeth Davidson
  - Mitchell Abrams

- Funding: NEH (HAA-261271-18, Digital Humanities Advancement Grant III A *Linked Digital Environment for Coptic Studies*)
What are entities?

Texts talk about things: people, places...

- Some of these are **named**:
  - “Emperor Diocletian”, “Alexandria”, ...

- But we also might care about **non-named** entities:
  - “A holy man”, “The desert”, “a different monastery”, ...
What are entities?

- For this talk: entities are **anything you can refer to** in a text (e.g. refer back to with a pronoun)
- Entities can be **nested** and belong to many **types**:

  - Τ ΠΙΣΤΙΣ ΣΩΣΙΝ Τ ΚΑΘΟΛΙΚΗ ΕΚΚΛΗΣΙΑ
  - also the faith of [the Catholic Church]

**Main question:** how can we **expose** these entities to users?
Goals

▪ Provide gold standard (non-)named entity data
▪ **Link** named entities to searchable stable identifiers
▪ Publish tools for **automatic** entity recognition
Data

- As a pilot project, we have manually annotated the Coptic Treebank (Zeldes & Abrams 2018) for entities:
  - 46,000 words
  - 6,500 entity mentions
    - 610 named (≈10%)
    - 441 identifiable, linked to 104 unique Wikipedia articles
Why Wikipedia?

- Common practice in Computational Linguistics – entity linking="Wikification" (Shnayderman et al. 2019)
- Offers a table of authorities out-of-the-box
- Very stable, (rare) identifier changes are tracked
- Comparable and linkable to other language resources
- Automatically gives us Wiki articles about mentioned entities, GIS (geo-location) information, …
 Wikification and links

- Users of Coptic Scriptorium data can immediately find out who someone/where some place is
- Get quantitative/distributional information
What’s included?

- Currently: Any nominal referring expression, except pronouns

- Ten entity classes:

<table>
<thead>
<tr>
<th>Cloud</th>
<th>ABSTRACT ('humility', 'thoughts')</th>
<th>Person</th>
<th>PERSON ('Ruth the Moabite', 'all the angels')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paw</td>
<td>ANIMAL ('a dog', '200 horses')</td>
<td>Place</td>
<td>PLACE ('a mountain', 'Alexandria')</td>
</tr>
<tr>
<td>Bell</td>
<td>EVENT ('his death', 'war')</td>
<td>Plant</td>
<td>PLANT ('the tree', 'fruit', 'wheat')</td>
</tr>
<tr>
<td>Box</td>
<td>OBJECT ('bottles', 'her hand')</td>
<td>Substance</td>
<td>SUBSTANCE ('the water', 'blood')</td>
</tr>
<tr>
<td>Castle</td>
<td>ORGANIZATION ('the king’s army', 'Catholic Church')</td>
<td>Time</td>
<td>TIME ('the month of Parmoute', 'ten years')</td>
</tr>
</tbody>
</table>
What can users search for?

- Entity types, identities
- Named/non-named

Since our data is also 100% syntactically parsed:

- Entity head noun ([the **army** of the emperor])
- Part of speech
- Grammatical function
- Foreign/Greek loan words
- ...

Exposing Coptic Entities
What can users search for?

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>entity -&gt; head pos=&quot;ART&quot;</td>
<td>entities headed by articles, ηετ...</td>
</tr>
<tr>
<td>text=/.<em>Π ΝΟΥΤΕ.</em>/</td>
<td>full text contains Π ΝΟΥΤΕ</td>
</tr>
<tr>
<td>entity=&quot;person&quot; <em>i</em> entity=&quot;place&quot;</td>
<td>PERSON containing a PLACE</td>
</tr>
<tr>
<td>identity=/.<em>Herod.</em>/</td>
<td>linked entity contains ‘Herod’</td>
</tr>
<tr>
<td>identity=&quot;Jesus&quot; &amp; meta::corpus=/.<em>mark.</em>/</td>
<td>mentions of Jesus in Mark</td>
</tr>
<tr>
<td>entity=&quot;person&quot; -&gt; head norm &amp; pos=&quot;V&quot; &amp; #3 -&gt; dep[func=&quot;obj&quot;] #2</td>
<td>verb with PERSON as object</td>
</tr>
<tr>
<td>entity=&quot;place&quot; <em>i</em> lang=&quot;Greek&quot;</td>
<td>PLACE containing Greek loanword</td>
</tr>
</tbody>
</table>

- **Try it here:**
  - [https://corpling.uis.georgetown.edu/annis/scripotorium/](https://corpling.uis.georgetown.edu/annis/scripotorium/)
  - Select *coptic.treebank*, enter a query from the above and search!

- **Detailed ANNIS QL reference:**
  - [https://corpus-tools.org/annis/aql.html](https://corpus-tools.org/annis/aql.html)
  - [https://copticscriptorium.org/ANNIS-tips.html](https://copticscriptorium.org/ANNIS-tips.html)
What can users search for?

βασκ υμπαρι αποκυρια, απεκασμα της Κωνσταντινούπολης, η οποία είναι γνωστή για την ιστορική σημασία της.

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?

What can users search for?
Automatic processing

- The Treebank is limited in scale
- Our goal is to include automatic entity annotations for all Coptic Scriptorium corpora by the end of summer
- Use Natural Language Processing tools

Approach:

- Rely on syntax parser to identify entity mention boundaries (also tested: Recurrent Neural Network)
- Use state-of-the-art machine learning classifiers and high coverage knowledge base to classify entities (using CRF classifier)
- Cascaded entity link matching – full text, entity head, corpus and frequency biases
Entity detection: trees and neural networks

**Figure 1.** Universal Dependencies tree for a Coptic sentence: “nor has the land ever adorned itself with a tomb” (Proclus, homily 13 ‘On Easter’, in Budge ed., urn:cts:copticLit:proclus.homily13.budge)

**Tree based detection**

**Recurrent Neural Network (RNN)**

Bi-LSTM with biaffine attention, Yu et al. (2020)
## Evaluation – entity detection

<table>
<thead>
<tr>
<th>method</th>
<th>Span match</th>
<th>Head match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>LOOKUP (baseline)</td>
<td>0.386</td>
<td>0.555</td>
</tr>
<tr>
<td>NOUN (entities = nouns)</td>
<td>0.123</td>
<td>0.111</td>
</tr>
<tr>
<td>TREE (gold parse)</td>
<td>0.879</td>
<td>0.862</td>
</tr>
<tr>
<td>TREE (pred parse)</td>
<td>0.831</td>
<td>0.815</td>
</tr>
<tr>
<td>RNN (binary)</td>
<td>0.653</td>
<td>0.732</td>
</tr>
</tbody>
</table>

**Span match:**

- oγ δωμε
- επίσχαν νογχη μυ
- πεπ σροσ

**Head match:**

- oγ δωμε
- επίσχαν νογχη μυ
- πεπ σροσ
## Evaluation – entity classification

<table>
<thead>
<tr>
<th>method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAJORITY</td>
<td>0.213</td>
<td>0.209</td>
<td>0.211</td>
<td>0.235</td>
<td>0.230</td>
<td>0.232</td>
</tr>
<tr>
<td>RNN</td>
<td>0.476</td>
<td>0.614</td>
<td>0.536</td>
<td>0.527</td>
<td>0.757</td>
<td>0.621</td>
</tr>
<tr>
<td>KB</td>
<td>0.681</td>
<td>0.660</td>
<td>0.670</td>
<td>0.728</td>
<td>0.705</td>
<td>0.717</td>
</tr>
<tr>
<td>CRF</td>
<td>0.805</td>
<td>0.778</td>
<td>0.791</td>
<td>0.861</td>
<td>0.831</td>
<td>0.846</td>
</tr>
<tr>
<td>CRF+KB</td>
<td><strong>0.827</strong></td>
<td><strong>0.810</strong></td>
<td><strong>0.818</strong></td>
<td><strong>0.889</strong></td>
<td><strong>0.869</strong></td>
<td><strong>0.879</strong></td>
</tr>
</tbody>
</table>

Knowledge Base

Conditional Random Fields (CRF)

91%

7%
Evaluation – entity linking

- **GitDox interface** (Zeldes & Zhang 2016)

- (corpus) > best match > (corpus) > best head

<table>
<thead>
<tr>
<th>method</th>
<th>accuracy</th>
<th>coverage</th>
<th>no error</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>0.227</td>
<td>0.273</td>
<td>0.953</td>
</tr>
<tr>
<td>head</td>
<td>0.433</td>
<td>0.500</td>
<td>0.933</td>
</tr>
<tr>
<td>cascade</td>
<td>0.460</td>
<td>0.500</td>
<td>0.960</td>
</tr>
</tbody>
</table>

**Entity types**

- **person:**
  
<table>
<thead>
<tr>
<th>Name</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ephrem</td>
<td>Ephrem the Syrian</td>
</tr>
<tr>
<td>Paul the Apostle</td>
<td></td>
</tr>
<tr>
<td>Saul</td>
<td></td>
</tr>
<tr>
<td>Jesus</td>
<td></td>
</tr>
<tr>
<td>Mary, mother of Jesus</td>
<td></td>
</tr>
<tr>
<td>Jesus</td>
<td></td>
</tr>
</tbody>
</table>
Metadata linking

- Enables search for documents by mentioned entities
Explore the data

- ‘distant reading’ interactive visualizations of the entity data:
  https://copticscriptorium.org/entities/breakdown.html
Automatic tools demo

- You can now use the Coptic-NLP tool suite to do automatic entity recognition!
- Online demo here, code and tools on GitHub:
  - [https://corpling.uis.georgetown.edu/coptic-nlp/](https://corpling.uis.georgetown.edu/coptic-nlp/)
Future plans

▪ Add automatic entity recognition to all Coptic Scriptorium corpora
▪ Manually correct named entity linking for core corpora (excl. OT, NT)
▪ Disseminate automatic entity tagging tools and data online
Mιωτήν τωνού!

Thank You
This work has been supported by

NEH Division of Preservation and Access (PW-51672-14)
NEH Office of Digital Humanities (HD-51907; HG-229371)
DFG
BMBF (German Federal Ministry of Education and Research)
Georgetown University
University of Oklahoma
The University of the Pacific
Humboldt University
Canisius College

Licensed under open licenses including

Apache 2.0 license, MIT license
Creative Commons 4.0 CC-BY license