New Techniques for Coding Political Events Across Languages

Yan Liang (yliang@ou.edu)

University of Oklahoma

Yan Liang, Andrew Halterman, Khaled Jabr, Christian Grant, Jill Irvine
Large - terabytes

Events Limited in English only

Who

Did

What

Event Coder

Language Specific Dictionaries

'I Would Make A Bad President,' Obama Says In Huge Campaign Blunder

Microsoft Ad Campaign Crashing Nation's Televisions

Supreme Court Upholds Stopping In The Name Of Love In 2-1 Decision
Coding Teams

• In order to assist with our dictionary development, we hired 8-10 Arabic coders.

• The coders were mostly undergraduate students and native Arabic speakers with direct experience in teaching the language.

• Coders were paired into groups of two with one performing a task and the second verifying.
Political Event Data

A “triple” of information:

an event such as an attack or protest, performed by a source actor, against a target.

"Turkey uses car bomb to attack Iraq."

<table>
<thead>
<tr>
<th>event</th>
<th>attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Turkey</td>
</tr>
<tr>
<td>target</td>
<td>Iraq</td>
</tr>
</tbody>
</table>
Dictionary Development

Resolving nouns (actors) and verbs (events) to common codes makes further analysis feasible.

Example:

- “demonstrated” and “rallied in the streets” would both be coded as 145:Protest violently, riot, not specified
- “Angela Merkel” and “German Ministry of Defense”, would be coded as DEU GOV
Solutions:

- CoreNLP-based interface
- NER-based interface
- Wiki-based interface
- Directed Translation.
Regular Coding Interface

- **Actor Coding**
- **Parsed Nouns**
- **Parsed Verbs**
- **Query Keyword**
- **LDA filtered topic**
- **Word2Vec derived synonym**
- **Not Sure Flag**
- **Verb coding**
Problems:

- CoreNLP parsing only consider grammar structure, so a lot of nouns and verbs might not be political event related.  
  Solution: NER-based interface

- Each actor might serve different roles at different times, that information is important when detecting new political event, coders spend a lot of time on those
  Solution: Wiki-based interface; prefill the role information
NER-based Interface

Five sentences contain the entity
Problems:

• The NER model trained in spaCy with "poor" data, so its performance is inadequate in recognizing person and organization names.

We tried to label more Arabic LOC, PER, ORG data
Wiki-based interface

Wiki link provided

Role name card prefilled
Problems:

- Not all politically relevant actors have Wikipedia pages,
- Nor do these pages always have biographical sidebars.
- Organizations also do not have biographical sidebars as people do.
Directed Translation method with no interface

Using this method we are able to get 5696 records in several hours.
Handle un-confidence coding:

The sentence that contains the actor at coding time displayed to give the content.

<table>
<thead>
<tr>
<th>SentenceId(click)</th>
<th>Word</th>
<th>Country</th>
<th>1st Role</th>
<th>2nd Role</th>
<th>Start(mm/dd/yyyy)</th>
<th>End(mm/dd/yyyy)</th>
<th>Confidence</th>
<th>AddBy</th>
<th>EditBy</th>
<th>AddTime</th>
<th>EditTime</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>58e04cab06036312a1edb21e</td>
<td>ألعاب</td>
<td>MIL</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0</td>
<td>Callin</td>
<td>yan1</td>
<td>Wed Apr 25 2018 13:00:43 GMT-0500 (CDT)</td>
<td>Sun Apr 29 2018 11:10:33 GMT-0500 (CDT)</td>
<td>Edit</td>
</tr>
<tr>
<td>58e045ce06036312a1e23a92</td>
<td>جزء الفعل الثنائي</td>
<td>IRQ</td>
<td>PTY</td>
<td>na</td>
<td>na</td>
<td></td>
<td></td>
<td>Callin</td>
<td>Wed Apr 25 2018 12:54:36 GMT-0500 (CDT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58e016ae06036312a1e60638</td>
<td>فرص</td>
<td>SAU</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td></td>
<td>Callin</td>
<td>Tue Apr 24 2018 11:48:11 GMT-0500 (CDT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58e01c6306036312a1ed96b8</td>
<td>عدم السؤال</td>
<td>SYR</td>
<td>MIL</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td></td>
<td>Callin</td>
<td>Tue Apr 24 2018 10:29:41 GMT-0500 (CDT)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance for each method

![Bar chart showing the number of actors coded by different coding approaches](image)

Fig. 6. Total number of actors coded for each approach
Discussion of coding speed

- The longer a coder has been coding overtime, and presumably the more experienced a coder becomes, the less average time it takes the coder to code an actor.
Summary:

- We were able to complete Arabic actor and verb dictionaries with coverage equivalent to English language dictionaries in less than two years of work compared to two decades that the English language dictionaries took to produce.
- We have use EventCoder to generate events from our corpus of millions of Arabic sources using the dictionary we developed, and we expect to make comparisons between it and the English corpus after final debugging and quality checking.
Future work:

• Use crowd sourcing on Wiki-based and NER-based coding to recommend action to coders.

E.g. we could make recommendations to our coders and ask them verify them instead of letting them enter detailed information. Prodigy is a promising framework that can provide us that functionality.
Future work:

• Enhance Arabic NER model.
  • Data:
    • OntoNotes Release 5.0
    • ANERCORP Data
    • Prodigy labelled data by our coders
  • Training Process
    • Spacy trained merged OntoNotes 5.0+ ANERCORP
    • Change the data into prodigy format, then mixed in the prodigy labelled data,
    • Update the model in order to avoid the catastrophic issue in successive model training.
THANK YOU.

oudalab.github.io
Discussion of coding speed

- Wiki-based approach is unexpectedly slow. We expected it to be faster than the NER-based system since we had already pre-populated the time range for each entity and provided the URL to link the actor back to their Wikipedia page.

<table>
<thead>
<tr>
<th>Method</th>
<th>Actor Coded</th>
<th>skipped</th>
<th>Time each role(seconds)</th>
<th>Time each actor (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki-based</td>
<td>2459</td>
<td>NA</td>
<td>202</td>
<td>377</td>
</tr>
<tr>
<td>Ner-based</td>
<td>204</td>
<td>7180</td>
<td>NA</td>
<td>56</td>
</tr>
</tbody>
</table>
Prodigy Interface to label Arabic NER.

Fig. 8. Using Prodigy to train a named entity recognition system
Gold Standard event coding report: