creating an automated event data system for arabic text

Andy Halterman, Jill Irvine, Christan Grant, Khaled Jabr, Yan Liang

4 April 2018

motivation

- text is a key source of data for political scientists
- increasing use of automated text analysis
- most automated analysis is for English

event data

Who did what to whom:

> *"German Chancellor Angela Merkel criticized the Turkish government for its restrictions on free speech"*

Actor: German Chancellor Angela Merkel $\rightarrow$ DEU GOV
Target: Turkish government $\rightarrow$ TUR GOV
Event: "criticized. . . restrictions on speech" $\rightarrow$ CONDEMN

Using the CAMEO event ontology

ecosystem

Large ecosystem of tools for event data:

- Petrarch2 (event coder)
  https://github.com/openeventdata/petrarch2
- phoenix_pipeline (end-to-end event data)
  https://github.com/openeventdata/phoenix_pipeline
- Birdcage (faster, distributed pipeline)
  https://github.com/openeventdata/birdcage/
- Mordecai (text geoparsing)
  https://github.com/openeventdata/mordecai/

steps in making event data

> *"German Chancellor Angela Merkel criticized the Turkish government for its restrictions on free speech"*

1. grammatical parsing of the sentence
2. find actor and event text
3. compare to dictionaries

Steps 1 and 2 are easy to change across languages. Step 3 requires unique dictionaries for each language.

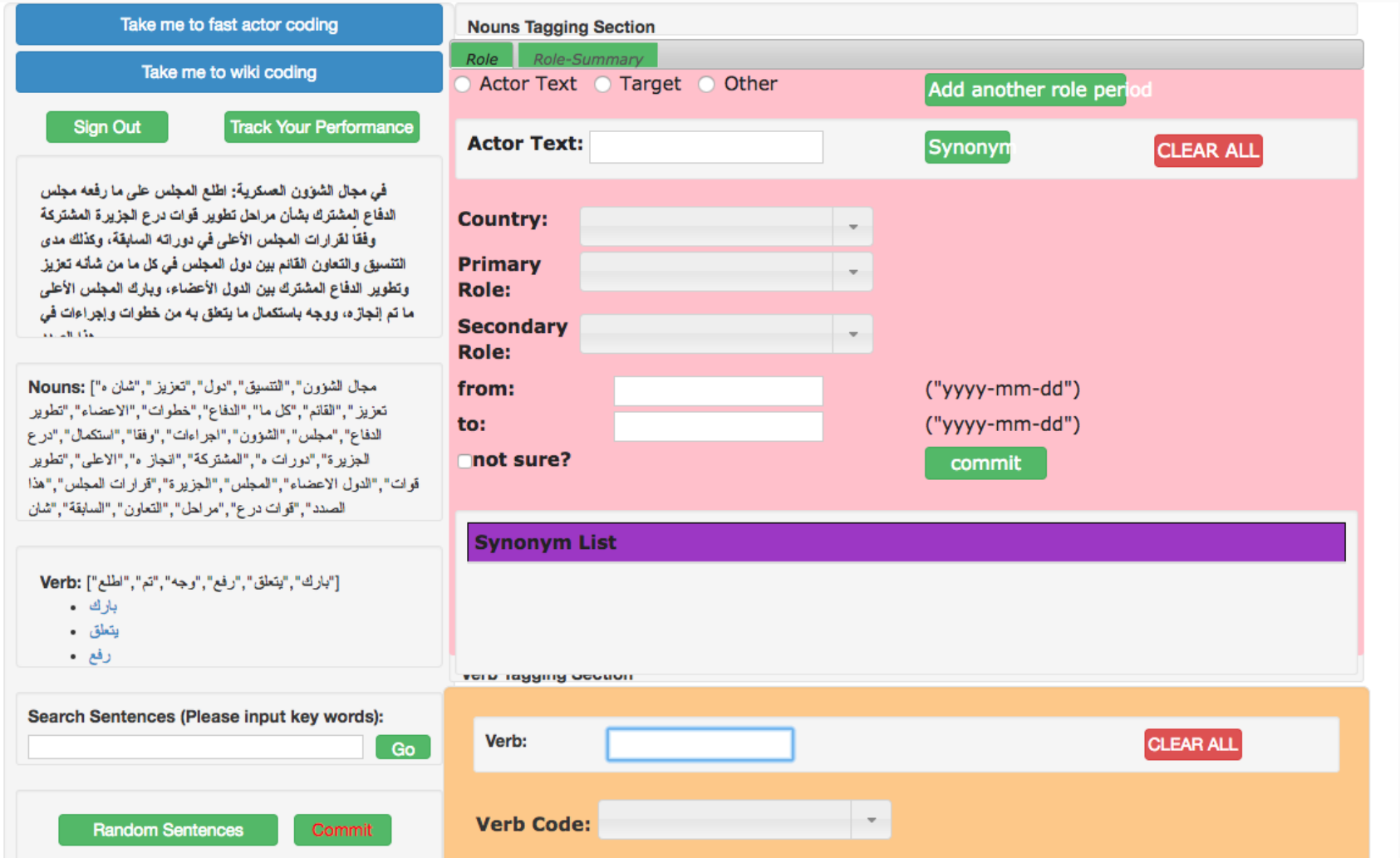
the role of dictionaries

The main bottleneck in making custom event data is in customizing dicitonaries.

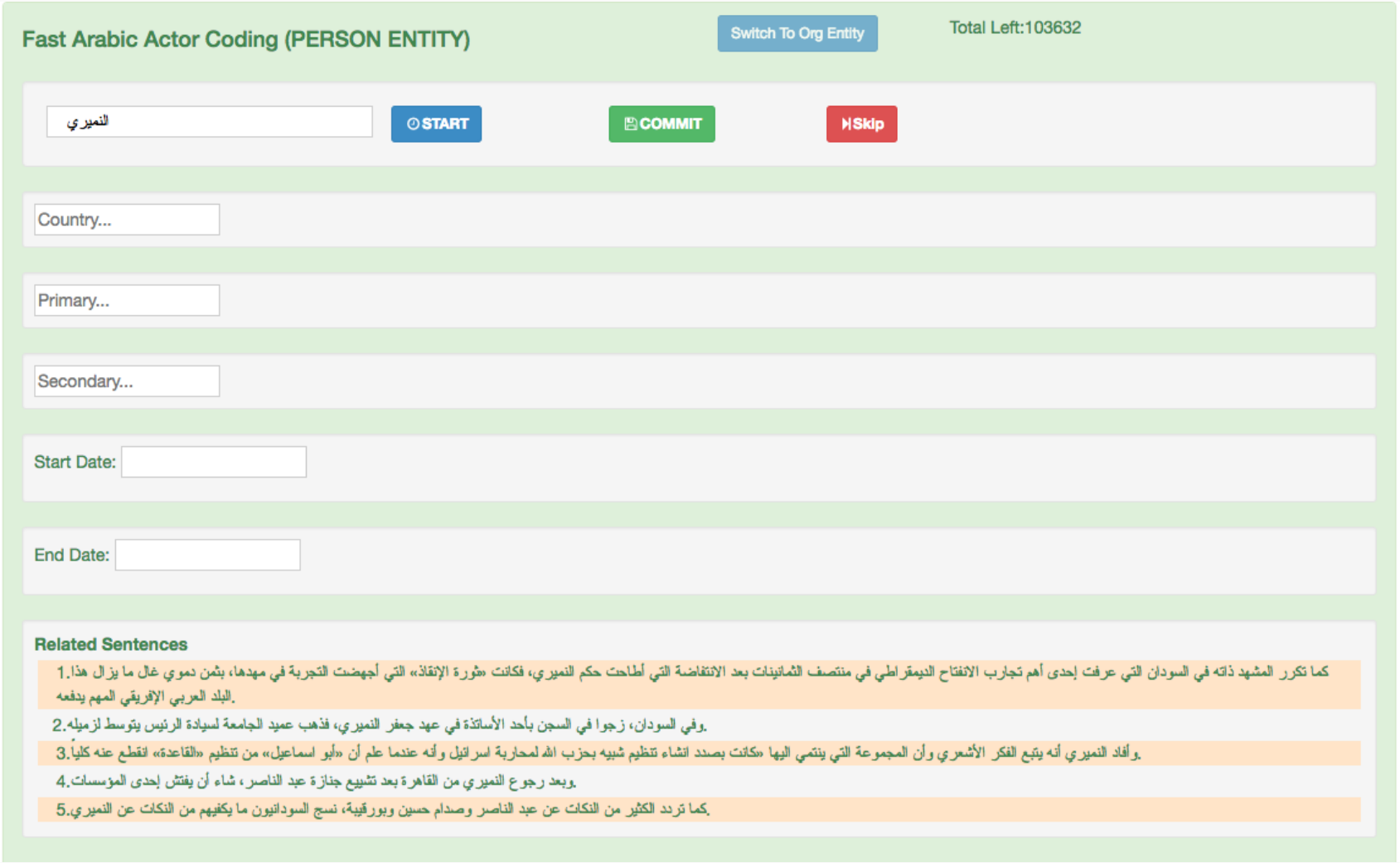
Understanding the best way to create dictionaries is useful for:

1. Making event data in other languages
2. Making dictionaries for new event types
3. Understanding where dictionaries come from

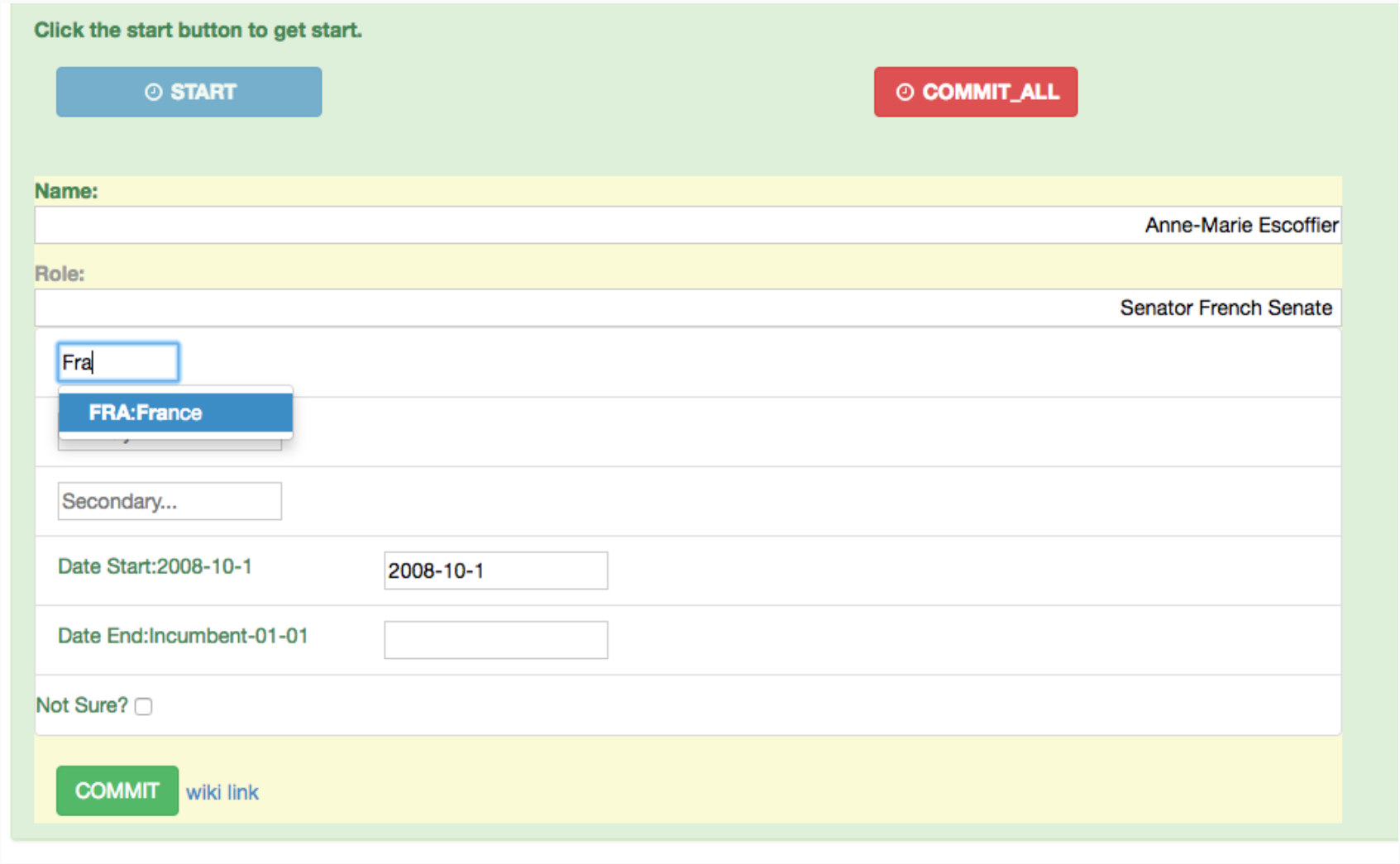
making actor dictionaries (text approach)



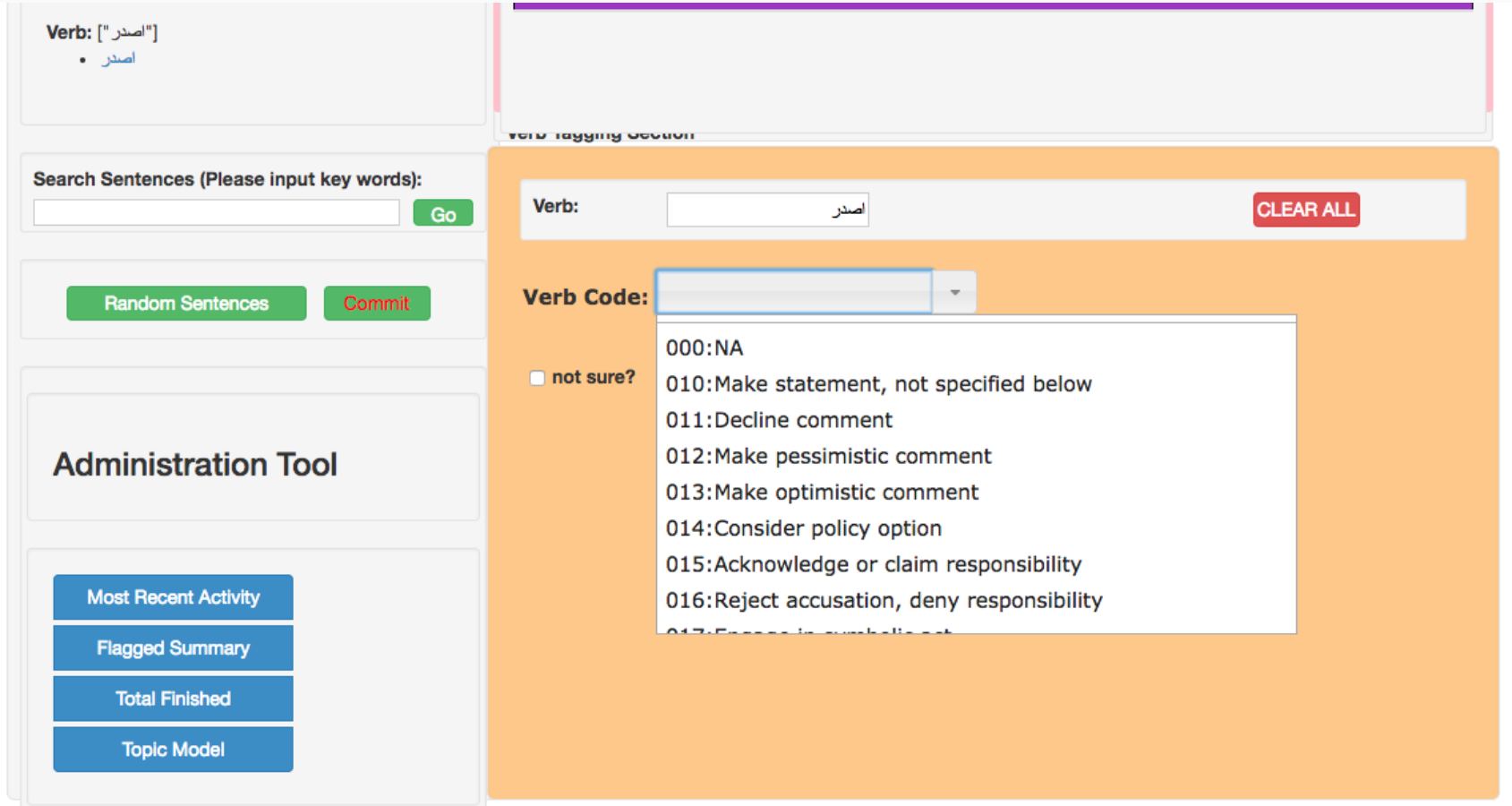
making actor dictionaries (ner approach)



making actor dictionaries (wiki approach)



verbs (text approach)



results

method	total actors coded	total verbs coded
regular interface	6,387	1,628
wiki translation	5,696	NA
NER coding	179 (with 6,667 skipped)	NA
wiki bio coding	2,327	NA
CAMEO translation	NA	~9,000

- Use Wikipedia where possible. Use raw text as a last resort
- Start by translating existing verb dictionaries, but understand limitations

ongoing work

- How different is data created from Arabic text from English text? What mistaken inferences might be drawn from relying on English language text?

ongoing work

- How different is data created from Arabic text from English text? What mistaken inferences might be drawn from relying on English language text?
- Pre-war political mobilization and violence against civilians in civil war (Balcells 2017). Do protests provide the same signal as election returns? Does the theory hold in Syria?