New Techniques for Coding Political Events Across Languages

Yan Liang (yliang@ou.edu)

University of Oklahoma

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

Large - terabytes

Events Limited in English only

Event Coder

Language Specific Dictionaries

Who

Did

What

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."

event	attack
source	Turkey
target	Iraq

Dictionary Development

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

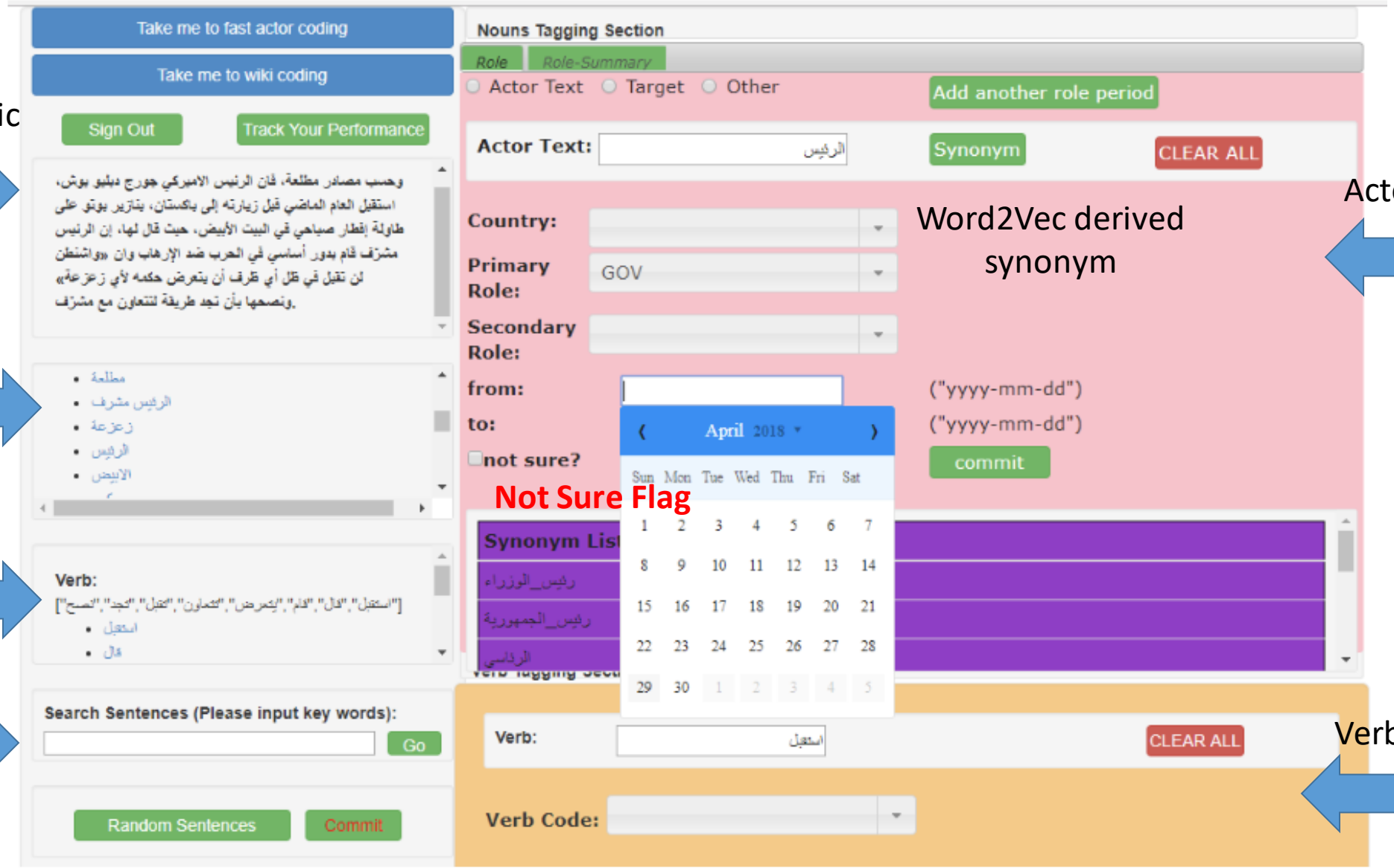
Example:

- “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.

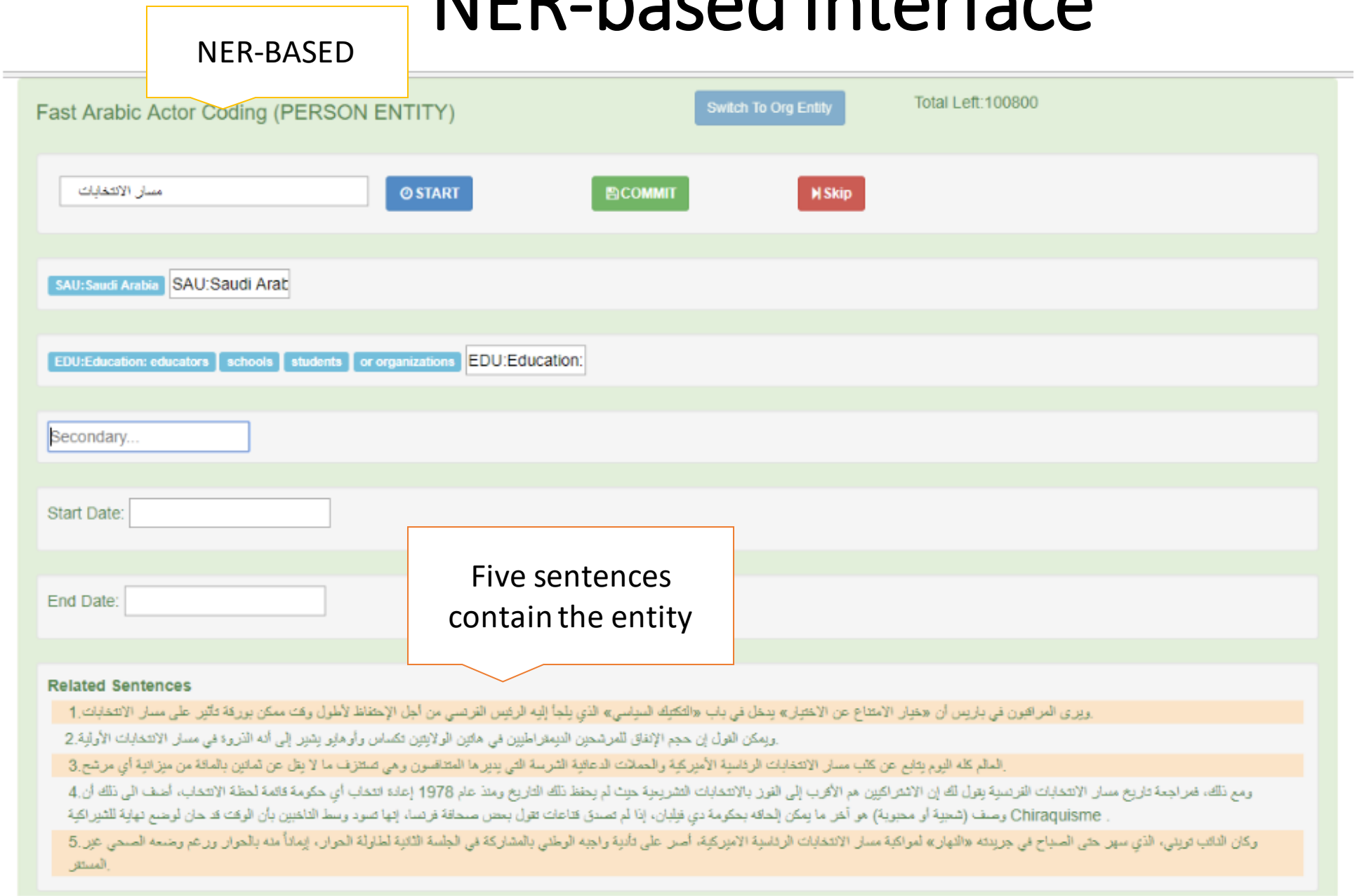
Regular Coding Interface



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

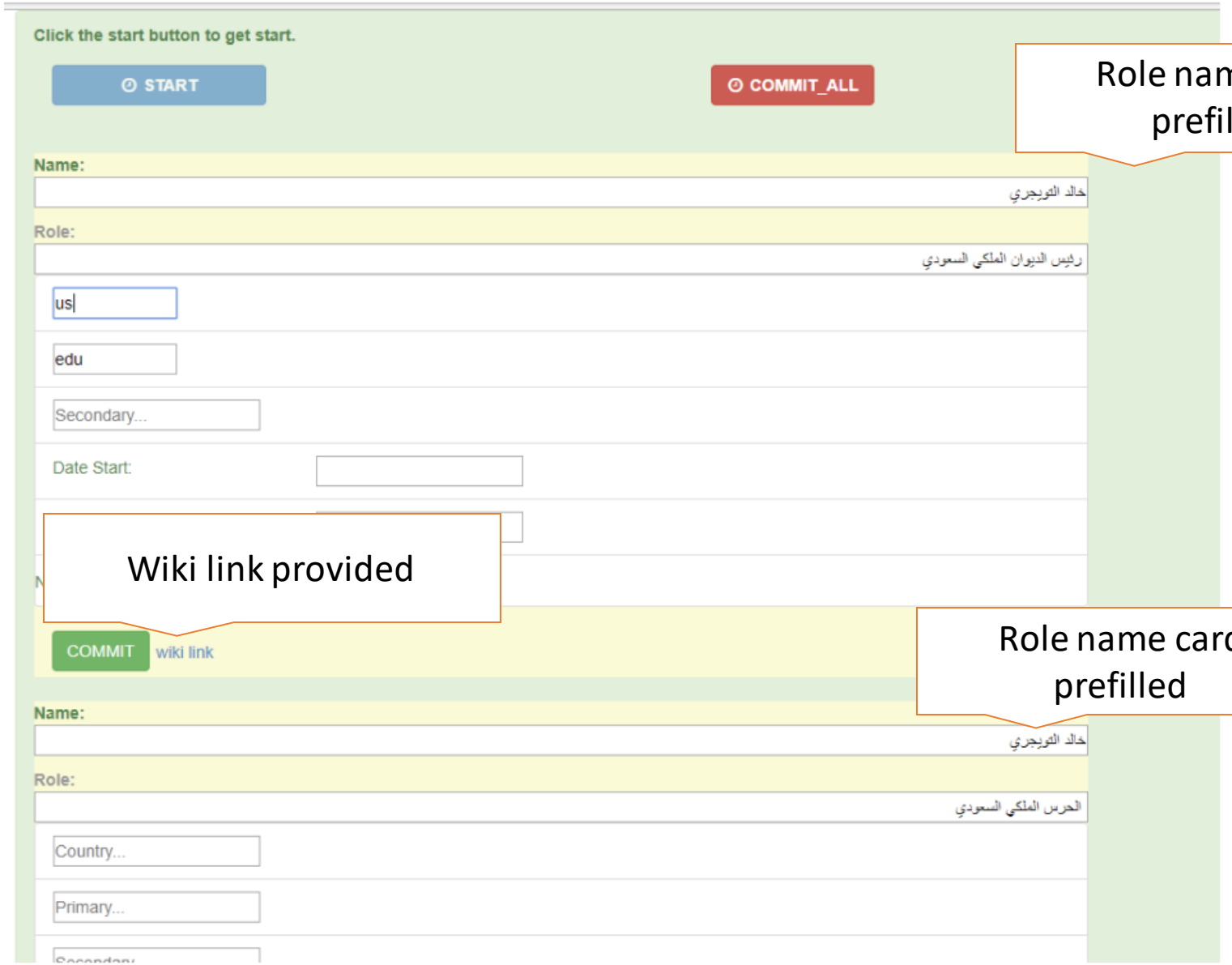


Problems:

- 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

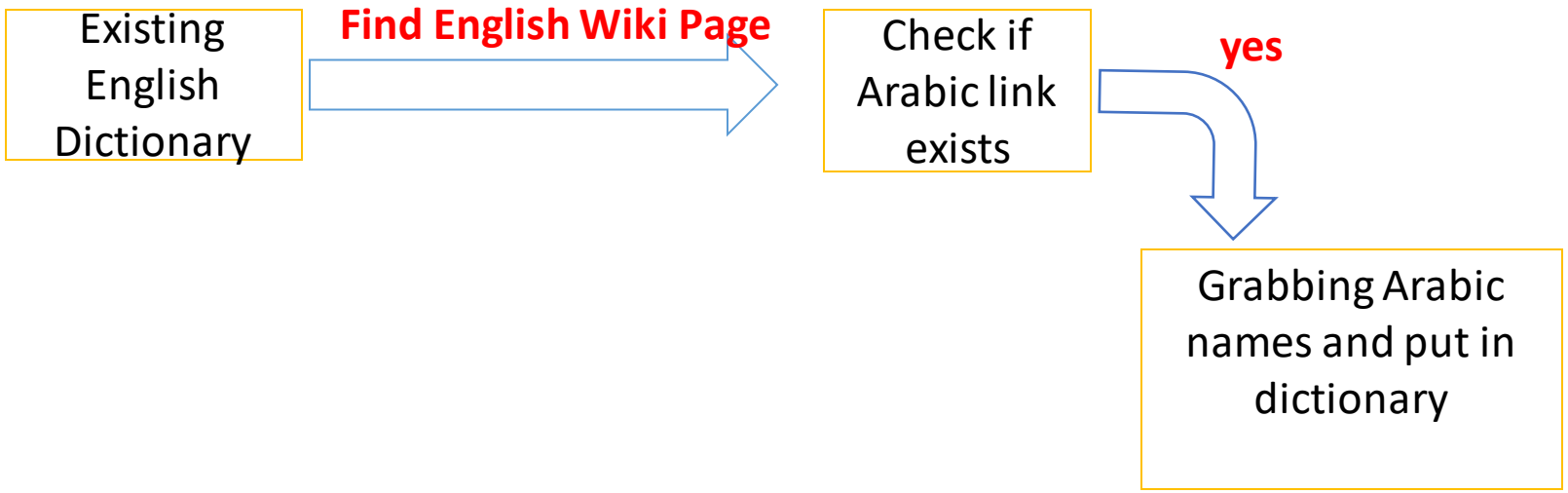
Wiki-based interface



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.

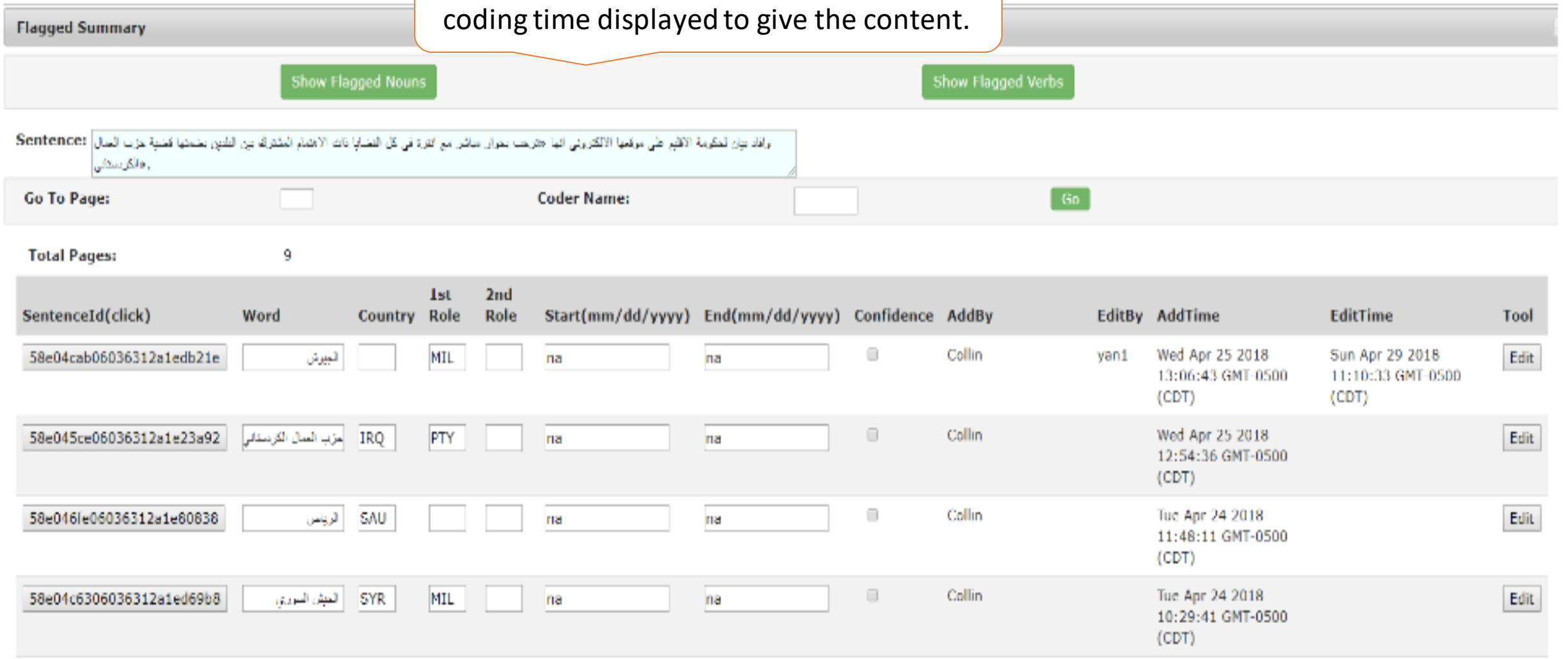
Directed Translation method with no interface



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.



Flagged Summary

Show Flagged Nouns | Show Flagged Verbs

Sentence: وافاد بيان لحكومة الاقليم على موقعها الالكتروني انها «ترحب بحوار مباشر مع اربيل في كل القضايا ذات الاهتمام المشترك بين الطرفين بضمنها قضية حزب العمال والكردستاني».

Go To Page: | Coder Name: | Go

Total Pages: 9

SentenceId(click)	Word	Country	1st Role	2nd Role	Start(mm/dd/yyyy)	End(mm/dd/yyyy)	Confidence	AddBy	EditBy	AddTime	EditTime	Tool
58e04cab06036312a1edb21e	الجيش		MIL		na	na		Collin	yan1	Wed Apr 25 2018 13:06:43 GMT-0500 (CDT)	Sun Apr 29 2018 11:10:33 GMT-0500 (CDT)	Edit
58e045ce06036312a1e23a92	حزب العمال الكردستاني	IRQ	PTY		na	na		Collin		Wed Apr 25 2018 12:54:36 GMT-0500 (CDT)		Edit
58e016fe06036312a1e80838	الرئيس	SAU			na	na		Collin		Tue Apr 24 2018 11:48:11 GMT-0500 (CDT)		Edit
58e01c6306036312a1ed69b8	الجيش السوري	SYR	MIL		na	na		Collin		Tue Apr 24 2018 10:29:41 GMT-0500 (CDT)		Edit

Performance for each method

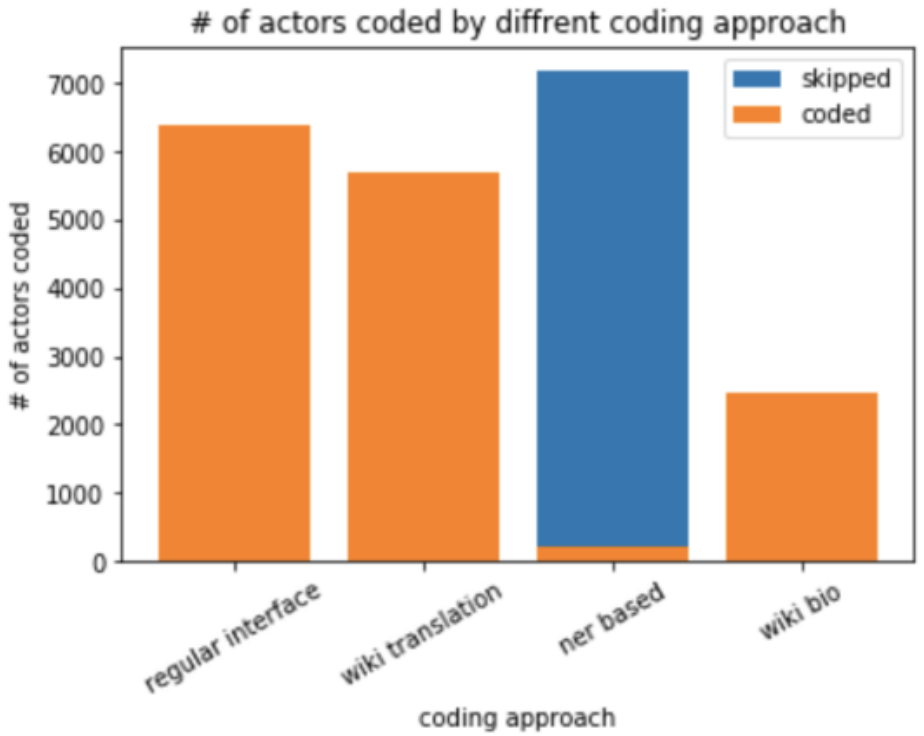
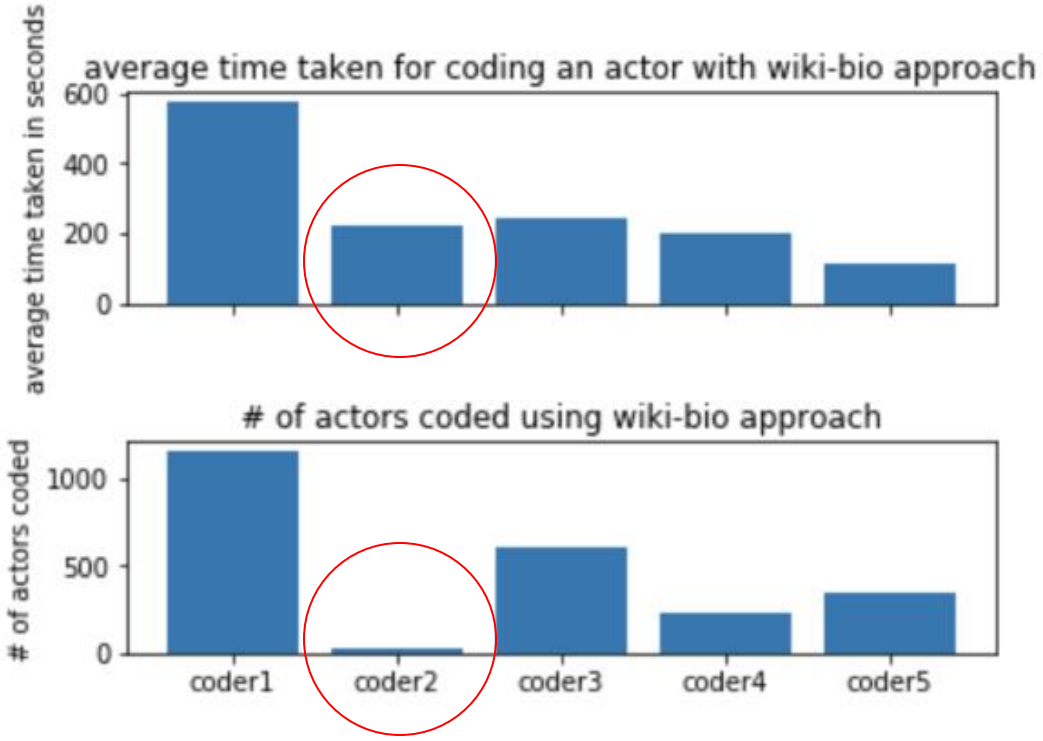


Fig. 6. Total number of actors coded for each approach

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.



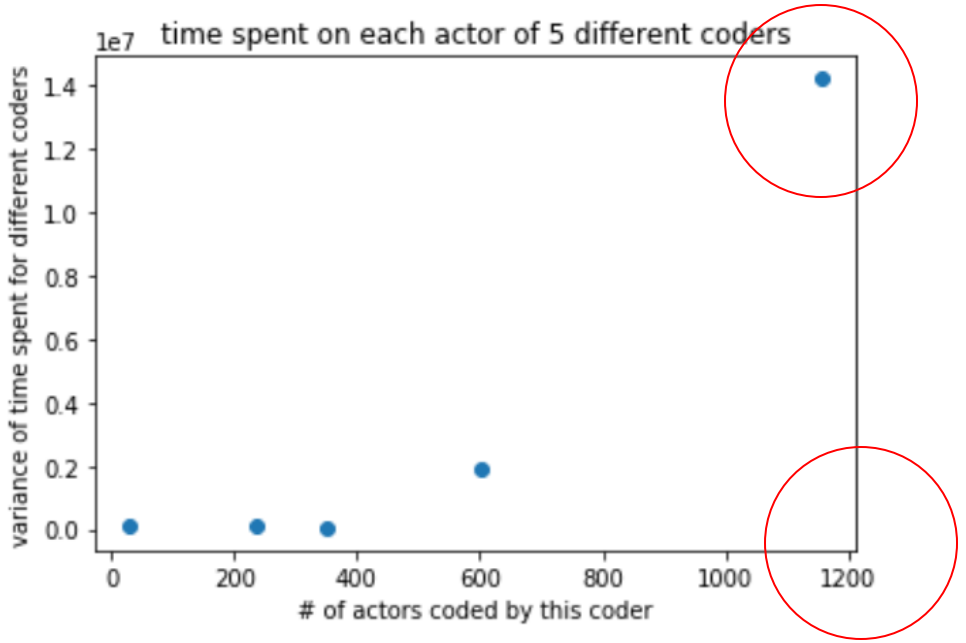


Fig. 7. Variance of time spent on each actor of different coders versus the number of actors coded by each coder

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.

Birdcage

Basic, Integrated, and Reliably Distributed
Coding, Actors, and Geolocation for Events

THANK YOU.

oudalab.github.io

Terrier

Temporally Extended Regularly
Reproducible International Event Records

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.

Method	Actor Coded	skipped	Time each role(seconds)	Time each actor (seconds)
Wiki-based	2459	NA	202	377
Ner-based	204	7180	NA	56

Prodigy Interface to label Arabic NER.

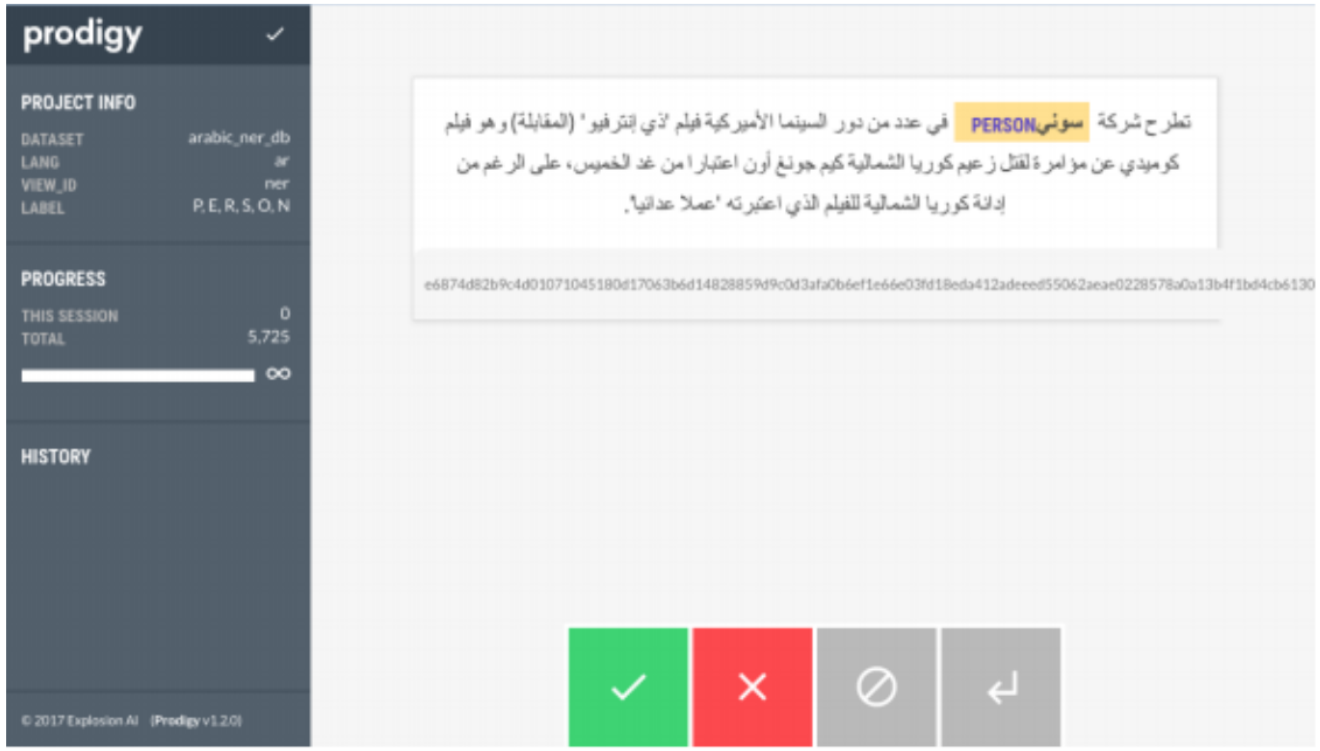


Fig. 8. Using Prodigy to train a named entity recognition system

Gold Standard event coding report: