

# 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

The interface is divided into several functional areas:

- Navigation:** Buttons for "Take me to fast actor coding" and "Take me to wiki coding".
- User Actions:** "Sign Out" and "Track Your Performance" buttons.
- Text Input:** "Actor Text:" field containing the Arabic word "الرئيس".
- Role Tagging Section:** Includes "Role" and "Role-Summary" tabs, radio buttons for "Actor Text", "Target", and "Other", and an "Add another role period" button.
- Metadata Fields:** "Country:", "Primary Role:" (set to "GOV"), and "Secondary Role:" dropdowns.
- Date Range:** "from:" and "to:" fields with a calendar pop-up for April 2018.
- Verb Tagging Section:** "Verb:" field containing "استقبل" and a "Verb Code:" dropdown.
- Buttons:** "Synonym", "CLEAR ALL", and "commit" buttons.

LDA filtered topic



Parsed Nouns



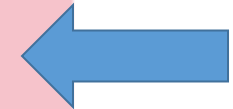
Parsed Verbs



Query Keyword



Actor Coding



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

NER-BASED

Fast Arabic Actor Coding (PERSON ENTITY) Switch To Org Entity Total Left: 100800

START COMMIT Skip

SAU: Saudi Arabia

EDU: Education: educators schools students or organizations

Start Date:

End Date:

Five sentences contain the entity

**Related Sentences**

1. ويرى المراهبون في باريس أن «خيار الامتناع عن الاختيار» يدخل في باب «التكتيك السياسي» الذي يلجأ إليه الرئيس الفرنسي من أجل الاحتفاظ لأطول وقت ممكن بورقة تأثير على مسار الانتخابات.
2. ويمكن القول إن حجم الإنفاق للمرشحين الديمقراطيين في هاتين الولايتين تكساس وأوهايو يشير إلى أنه الذروة في مسار الانتخابات الأولية.
3. العالم كله اليوم يتابع عن كثب مسار الانتخابات الرئاسية الأميركية والحملات الدعائية الشرسة التي يديرها المتنافسون وهي صلتزف ما لا يقل عن ثمانين بالمائة من ميزانية أي مرشح.
4. ومع ذلك، فمراجعة تاريخ مسار الانتخابات الفرنسية يقول لك إن الاثراكيين هم الأقرب إلى الفوز بالانتخابات التشريعية حيث لم يحفظ تلك التاريخ ومدت عام 1978 إعادته انتخاب أي حكومة قائمة لحظة الانتخاب، أضف الى ذلك أن Chiraquisme وصف (شجيرة أو محبوبة) هو آخر ما يمكن إلحاقه بحكومة دي فيليان، إننا لم تصدق كتاعات تقول بعض صحافة فرنسا، إنها تسود وسط الناخبين بأن الوقت قد حان لوضع نهاية للشيراكية وكان النائب تويدي، الذي سهر حتى الصباح في جريدته «النهار» لمواكبة مسار الانتخابات الرئاسية الأميركية، أصر على تأدية واجبه الوطني بالمشاركة في الجلسة الثابتة لمناقشة الحوار، إيماناً منه بالحوار ورغم وضعه الصحي غير المستقر.

# 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

Click the start button to get start.

**Name:**

**Role:**

Date Start:

[wiki link](#)

**Name:**

**Role:**

Role name card prefilled

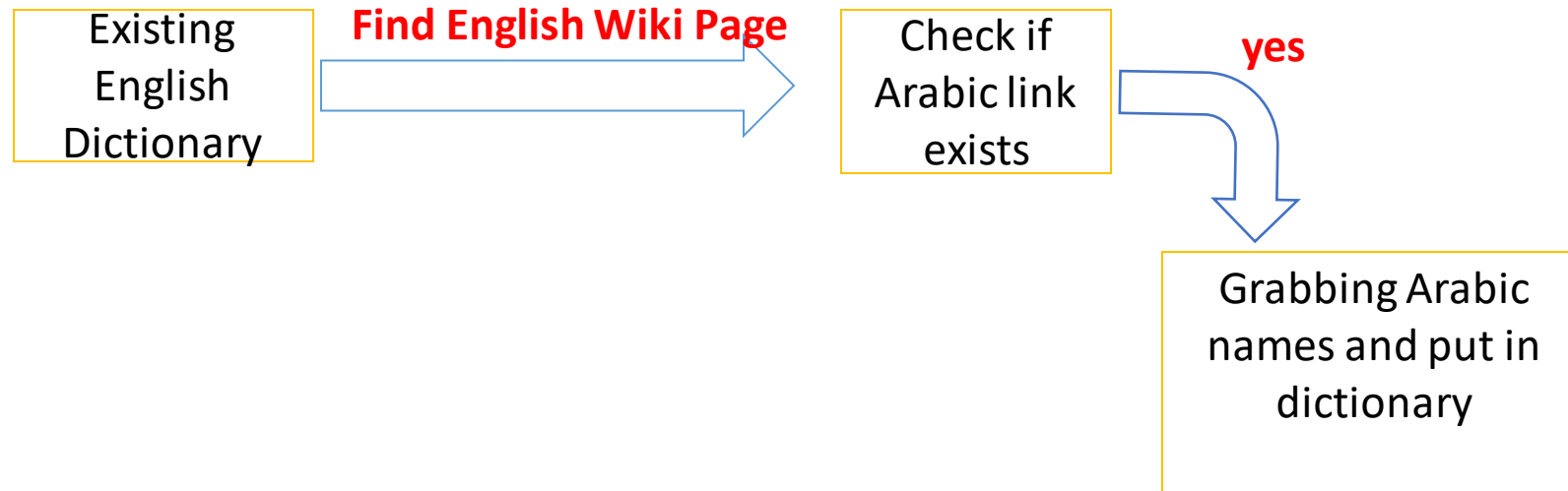
Wiki link provided

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.

Flagged Summary

Show Flagged Nouns

Show Flagged Verbs

Sentence:

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
<a href="#">58e04cab06036312a1edb21e</a>	<input type="text" value="الجيش"/>	<input type="text"/>	MIL	<input type="text"/>	<input type="text" value="na"/>	<input type="text" value="na"/>	<input type="checkbox"/>	Collin	yan1	Wed Apr 25 2018 13:06:43 GMT-0500 (CDT)	Sun Apr 29 2018 11:10:33 GMT-0500 (CDT)	<a href="#">Edit</a>
<a href="#">58e045ce06036312a1e23a92</a>	<input type="text" value="حزب العمال الكردستاني"/>	IRQ	PTY	<input type="text"/>	<input type="text" value="na"/>	<input type="text" value="na"/>	<input type="checkbox"/>	Collin		Wed Apr 25 2018 12:54:36 GMT-0500 (CDT)		<a href="#">Edit</a>
<a href="#">58e0461e06036312a1e80838</a>	<input type="text" value="الرياض"/>	SAU	<input type="text"/>	<input type="text"/>	<input type="text" value="na"/>	<input type="text" value="na"/>	<input type="checkbox"/>	Collin		Tue Apr 24 2018 11:48:11 GMT-0500 (CDT)		<a href="#">Edit</a>
<a href="#">58e04c6306036312a1ed69b8</a>	<input type="text" value="الجيش السوري"/>	SYR	MIL	<input type="text"/>	<input type="text" value="na"/>	<input type="text" value="na"/>	<input type="checkbox"/>	Collin		Tue Apr 24 2018 10:29:41 GMT-0500 (CDT)		<a href="#">Edit</a>

# Performance for each method

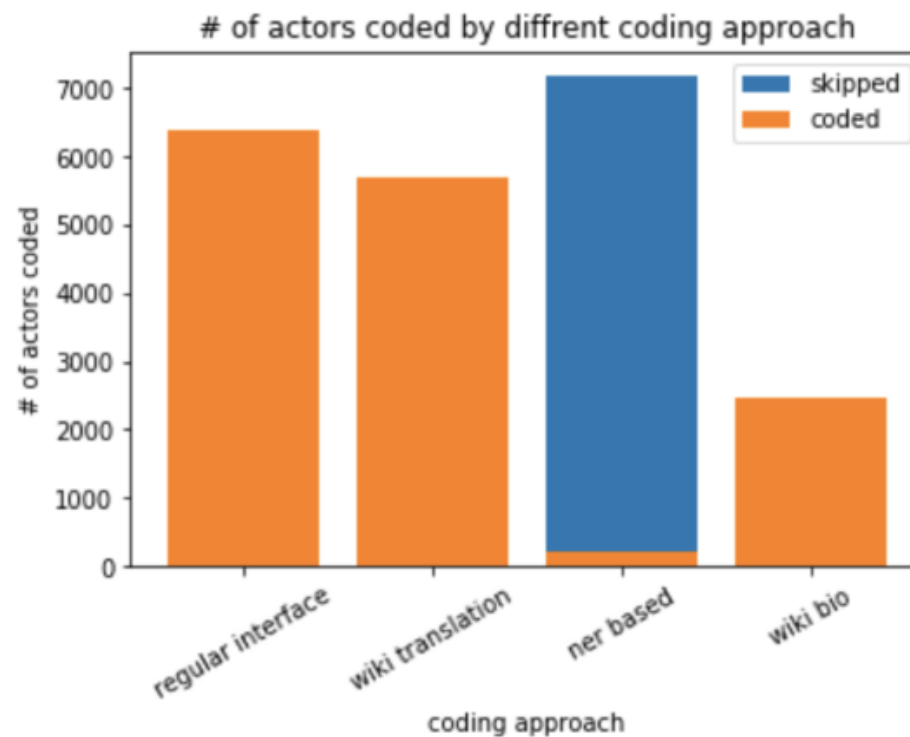


Fig. 6. Total number of actors coded for each approach

# 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.

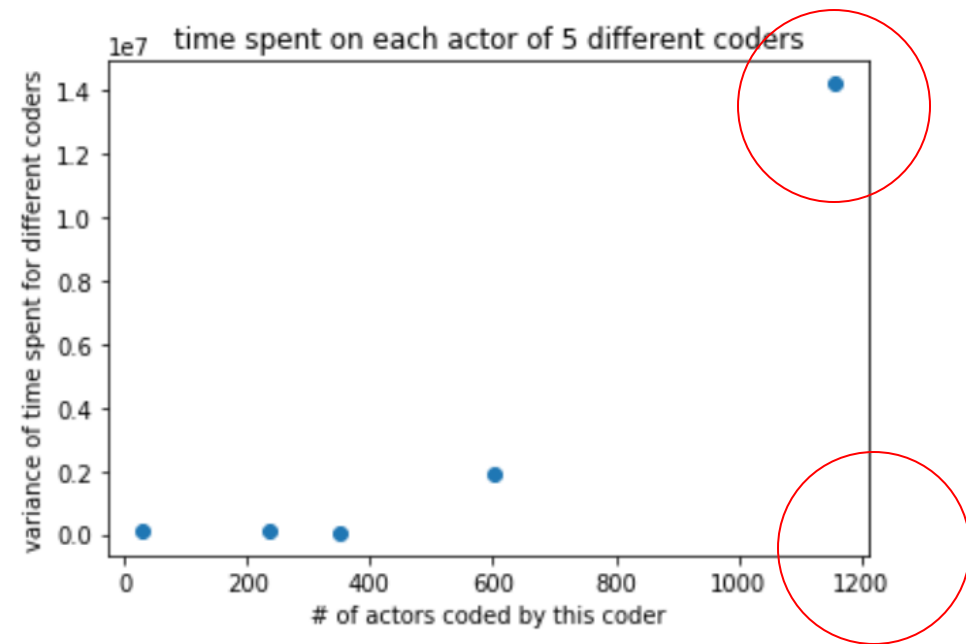
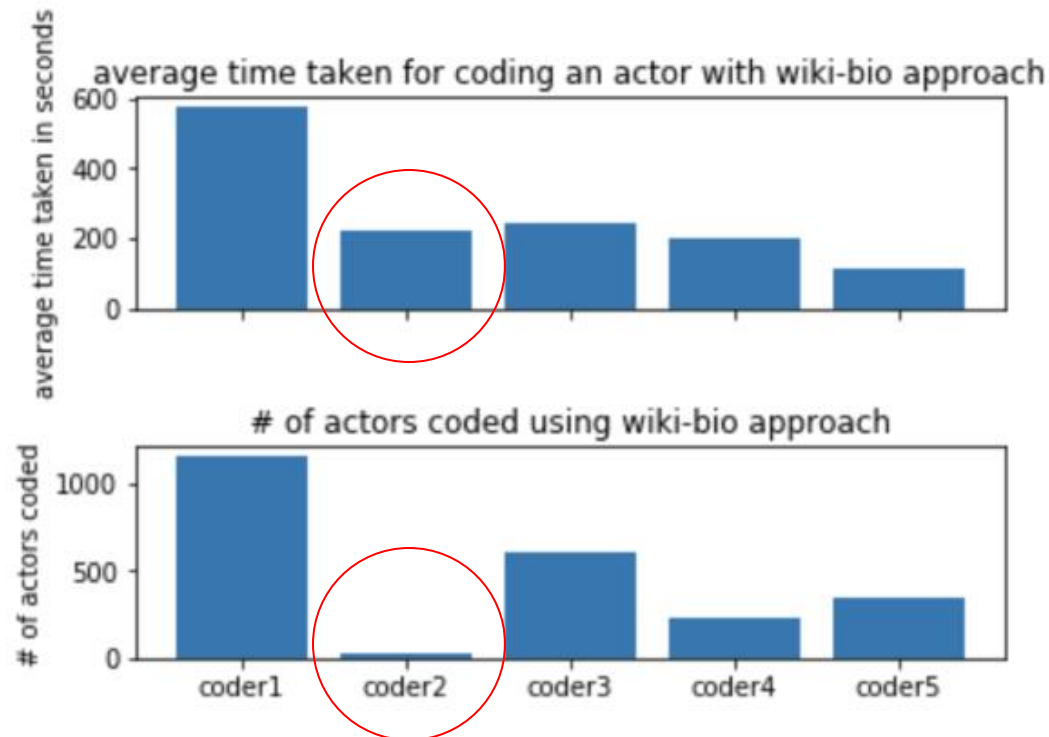


Fig. 7. Variance of time spent on each actor of different coders versus the number of actors coded by each coder



# 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.



THANK YOU.

[oudalab.github.io](https://oudalab.github.io)



## Birdcage

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



## 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:

