Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks

Andy Halterman (Michigan State), Philip A. Schrodt (Parus Analytics), Andreas Beger (Basil Analytics), Benjamin E. Bagozzi (Delaware), Grace I. Scarborough (Leidos)

International Studies Association, Montreal 2023

*This research was sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency. The views expressed in this paper are the authors' alone and do not represent the views of the U.S. Government. Funding for PLOVER was initially provided in part by the U.S. National Science Foundation award SBE-1539302.

Two Goals

We had two goals in making a new coder:

1. To produce new, high quality global event dataset (POLECAT) using the PLOVER ontology.
2. To develop tools for researchers to make custom, non-PLOVER datasets.

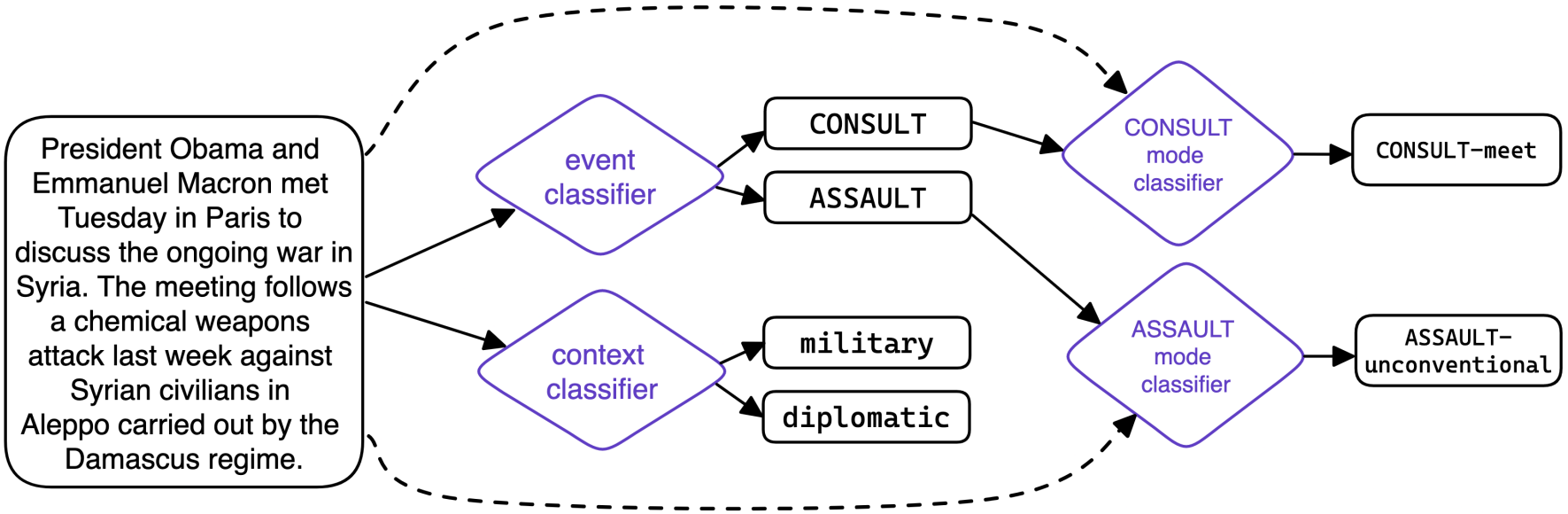
The “bag of tricks" we introduce dramatically lower the costs of developing new event coders and datasets.

Steps in event extraction

We conceptualize event extraction as containing the following 6 steps:

1. Event classification
2. Sub-event (“mode”) classification
3. Context classification
4. Event attribute identification
5. Actor, location, and date resolution
6. Entity categorization

Event, Mode, and Context Classification

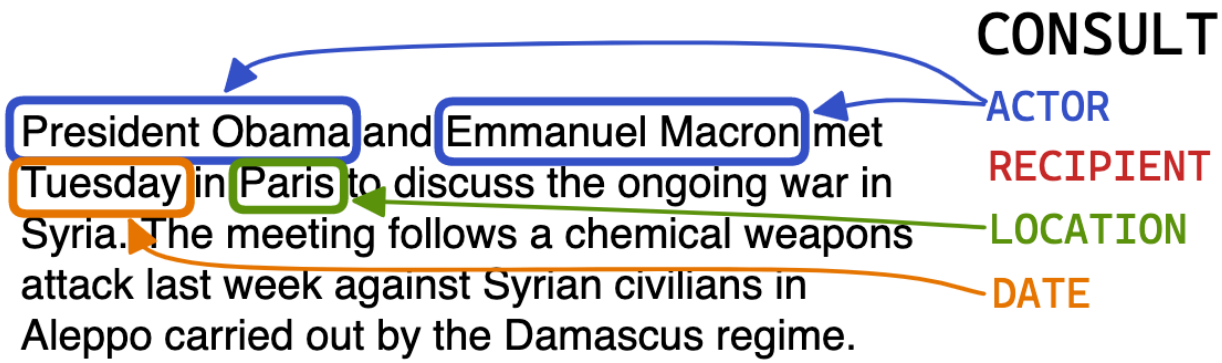


- We use document classifiers to identify events, modes, and contexts.
- We operate at the document instead of sentence level.
- We annotate several thousand documents with the PLOVER ontology and train transformer and SVM classifiers.

Event Attributes

An event's attributes correspond with the "who", "to whom", "when", and "where" questions.

We want to identify the answer to each question in the document.



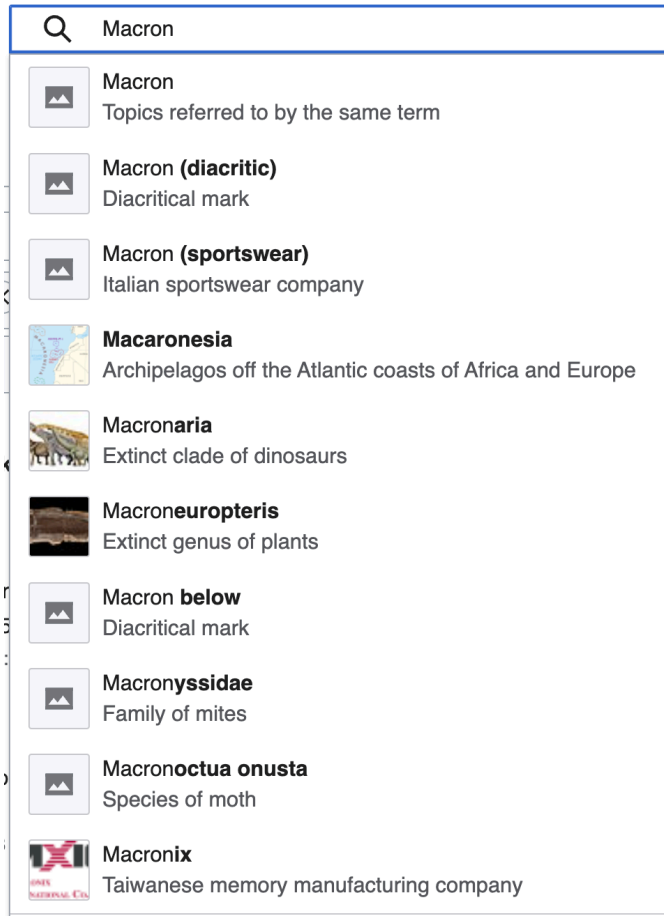
Identifying Attributes with a QA model

- We use extractive question answering to identify event attributes.
- We fine-tune a pretrained QA model on 3,000 newly annotated question+answer documents.
- We use event-mode specific questions for each attribute:

Event	Attribute	Question
PROTEST-demo	ACTOR	"Who held a demonstration?"
	ACTOR	"Who held a demonstration against {recip text}?"
	RECIP	"Who was the target of the demonstration?"
	RECIP	"Who was {actor text}'s demonstration against?"
	LOCATION	"Where was the demonstration held?"
PROTEST-riot	ACTOR	"Who engaged in the riot?"
	ACTOR	"Who rioted against {recip text}?"
	RECIP	"Who was the riot directed against?"
	RECIP	"Who did {actor text} riot against?"
	LOCATION	"Where did the riot take place?"

Actor Resolution–Querying Wikipedia

- In contrast to earlier event coders, we resolve named entities to Wikipedia.
  - Gives us a canonical name
  - Provides information on sector/job/country
- Download a Wikipedia dump and load into Elasticsearch
- Parse redirects and alternative names
- Query Wikipedia and pick best match using a rule+ML ranker



Actor Resolution–Entity Categorization

identify current role

25th President of France

Incumbent

Assumed office

14 May 2017

Prime Minister Édouard Philippe

Jean Castex

Élisabeth Borne

Preceded by François Hollande

Minister of Economics, Industry and Digital Affairs

In office

26 August 2014 – 30 August 2016

Prime Minister Manuel Valls

Preceded by Arnaud Montebourg

Succeeded by Michel Sapin

Assistant Secretary-General of the Presidency

In office

15 May 2012 – 15 July 2014

President François Hollande

Preceded by Jean Castex

Succeeded by Laurence Boone

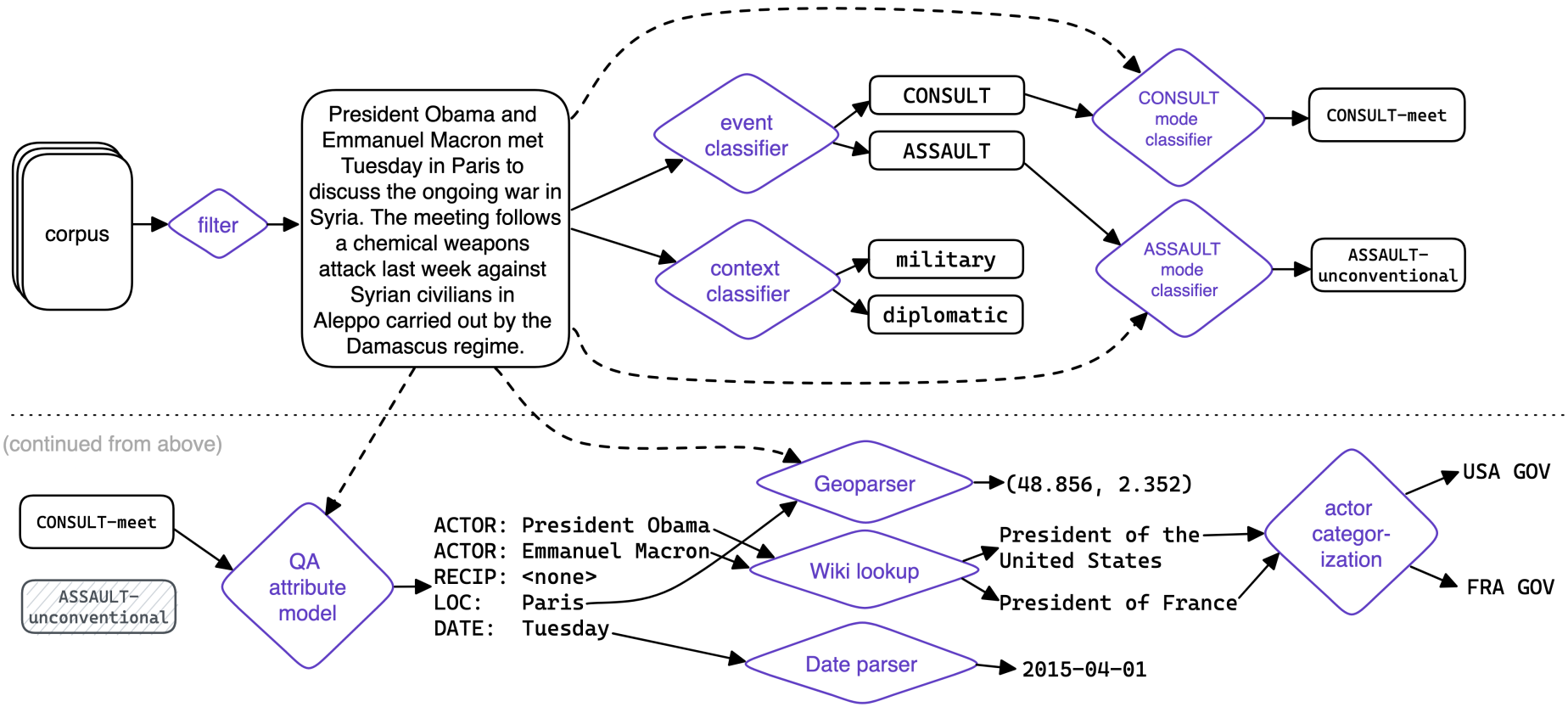
Compare the extracted office title to a list of generic titles and descriptions.

```
VILLAGE [CVL]
DEPARTMENT_OF_AGRICULTURE [GOVAGR]
DEFENSE_MINISTER [GOVMIL]
REBEL_LEADER [REB]
PRIME_MINISTER [GOV]
INTELLIGENCE_SERVICE [SPY]
```

We embed the search term and generic office titles using a neural model and find the entry with the closest cosine similarity.

Generic actors are compared directly to the file, skipping the Wikipedia lookup step.

Putting it all together...



How to make custom event data

1. Annotate new text with the desired event type/mode/context labels.
   - Active learning makes this process much more efficient.
   - Synthetic text addresses the rare class problem and may yield "zero-shot" classifiers.
2. Update QA model for domain and event types.
   - The Prodigy annotation interfaces makes it quick to collect QA spans.
3. Write actor $\rightarrow$ role code mapping file.
   - Write examples of which office titles/descriptions get mapped to which codes.

See the replication of the BFRS Pakistan political violence dataset in the paper.