RePAIR: Recommend Political Actors In Real-time From News Websites

Mohiuddin Solaimani, Sayeed Salam, and Latifur Khan
Department of Computer Science
The University of Texas at Dallas
Email: (mxs121731, sxs149331, lkhan)@utdallas.edu

***Abstract*—Extracting a structured representation of events (political, social etc.) has become an interesting domain in the computational and social sciences. A traditional approach is to use dictionary-based pattern lookups to identify actors and actions involved in potential events. A key complication of this approach is updating the dictionaries with new actors (e.g., when a new president takes office). Currently, the dictionaries are curated by humans, updated infrequently, and at a high cost. This means that tools dependent on the dictionaries (e.g., PETRARCH) overlook events because actors associated to them are missing in the dictionary. Moreover, these tools use only the syntactic structure of the sentence (e.g., parse tree, etc.) for their event coding. As a result, they generate events with inappropriate actors which fail to capture actual political interaction. To overcome these issues, we propose a framework *RePAIR* to recommend new political actors in real-time from the political news articles of different news websites (news articles with RSS feeds related to national/international politics) all over the world. The framework identifies semantic structure of a sentence using Automatic Content Extraction(ACE) method and uses a frequency based actor ranking algorithm and recommends most frequent new political actors over multiple time windows. We also suggest the associated role of recommended new actors from the role of co-occurred political actors in the existing CAMEO actor dictionary. Moreover, we integrate an external knowledge base (e.g., Wikipedia) into our framework to capture the evolving role of existing actors over time and recommend new roles for them. Furthermore, we consider PETRARCH and BBN ACCENT event coders for actor recommendation, and a graph-based actor role recommendation using weighted label propagation as baselines and compare them with our framework. Experimental results show our approaches outperform them significantly.**

***Index Terms*—PropBank, PETRARCH; SPARK; CAMEO; Actor Ranking**

I. Introduction

Political event data [1] are encoded from news reports. This can be accomplished manually by humans or via machine coding. Events are categorized based on a set of dictionaries for the actions and actors. The typical format to create these data is to determine "who-did/said-what to whom" [2]. While the set of actions or verbs is finite and can be matched to a fixed ontology, the set of source or target nouns is quite large in political terms. Identifying these political actors along with their roles is important in order to transform news reports into useful data for the study of international relations and civil conflict.

Currently, humans add new actors and their roles to the dictionary. Automated coders (i.e. PETRARCH [2]) use those dictionaries to identify events. However, if a source or target actor has *not* been entered by a human in the relevant actor dictionary, PETRARCH will ignore the event related to the new actor. So, we lose valuable information. To facilitate PETRARCH, we need automatic content extraction techniques, so that we can find possible new political actors from the events that PETRARCH ignores. This motivates us to design a framework for recommending new political actors in real-time. Furthermore, existing event coders (e.g., PETRARCH) use only the syntactic structure (i.e., parse tree) of a sentence which fails to capture the semantic of a sentence. More precisely, they struggle to encode event from a complex sentence (sentence with a prepositional clause, etc.). As a result, either they generate events with wrong actors (source/target) or fail to generate any event at all (more details in section II-B). This drives us to use the sentence semantic for our actor recommendation framework.

Automatic Content Extraction (ACE) programs [3] infer entities, relations, and events from human language data. An event consists of ACE relation which has a number of participants (ACE entities) and each participant has a semantic role that it plays in the event (agent, object, source, target). Machine learning-based semantic role labeling [4] technique consists of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. Example of tools implementing such technique includes FrameNet [5], PropBank [6], VerbNet [7]. Among them, PropBank Corpus provides structured predicate-argument annotation for the entire Penn Treebank [8]. Each verb in the treebank is annotated by a single instance in PropBank and acts like an event containing information about the source, target, and location of the verb.

Implementing such an automated system for recommending new political actors poses some key challenges. First, an actor may come with multiple alias names, e.g., 'Barack Hussein Obama', 'Barack Obama', 'President Obama', etc. Currently, a human expert puts a dictionary entry for each of these alias in CAMEO actor dictionary. Second, the role of an actor changes over time. For example, 'Shimon Peres' served multiple political roles in Israel during his lifetime. Finally, processing a large volume of news articles across the world in real-time demands a fast, scalable, and distributed computing solution.

Considering these challenges, we develop a real-time, distributed framework *RePAIR* to recommend new political actors

with their associated roles. The paper highlights the following contributions.

First, we have proposed a novel time window-based, unsupervised new actor recommendation technique with possible roles. In particular, we have designed a frequency-based actor ranking algorithm with alias actor grouping from news articles. We also integrate external knowledge-base (Wikipedia) to capture the timeline of an existing actor's role change and to suggest possible new roles if he/she has been assigned to a new political role.

Second, we have developed an ACE based semantic event coding instead of traditional syntactic event coding into our actor recommendation framework.

Third, we have developed a distributed real-time framework using Apache Spark Streaming [9] to address the scalability issue of event coding over political news articles.

Finally, we have compared our proposed approaches with state of the art (e.g., PETRARCH, BBN ACCENT event coding) and shown the effectiveness of our work.

The rest of the paper is organized as follows: Section **??** gives the background on the various open source tools and concepts used in this paper. Section II explains the overall framework. Section III describes our new actor with role recommendation technique in details. Section IV shows the experimental results for new actor detection. Section V covers the related works. Section VI states the conclusion and future works followed by a bibliography.

II. Framework

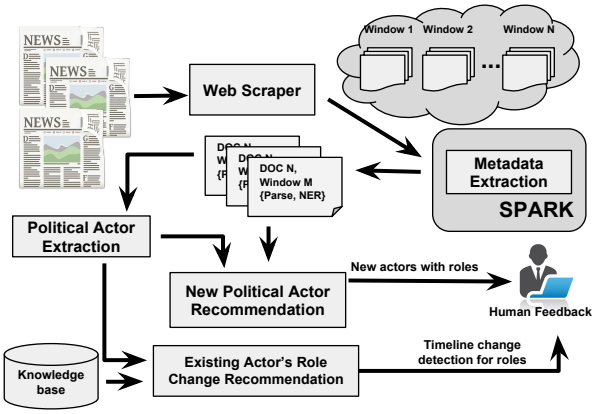


Fig. 1: Framework for real-time new political actor recommendation

Figure 1 shows our new political actor recommendation framework. *RePAIR* collects political news stories periodically through web scrapers. Later, it uses CoreNLP [10] and PropBank to extract metadata from the stories like parse trees, tokens, lemmas, NER, semantic events (by PropBank), etc. and stores them in MongoDB [11]. We use Apache Spark streaming [9] with Kafka [12] to collect all the scraped data, periodically. We use CoreNLP and PropBank annotation inside a Spark worker node to scale up the process. After that, it classifies political events only using CAMEO action dictionary from the metadata. It fetches possible new actors from source and target of the events and NER. We implement a window-based, unsupervised frequency-based actor ranking technique to recommend potential new actors with their related roles using event codes and associated metadata. Finally, a human expert validates the recommended actors with roles and updates the dictionary. We describe our framework in details in following subsections.

A. Web Scraper

We use a web scraper [13] to collect news articles and extract the content of the news. The web scraper has been integrated into the framework and runs periodically (e.g., every 2 hours). It collects news stories from 400 RSS (Rich Site Summary) Feeds[14] (on average, each contains 15-20 links to news articles). The news articles are shipped through Apache Kafka to an Apache Spark-based data processing module[15].

B. Metadata Extraction

Inside our data processing unit, CoreNLP parses all the news documents and extracts metadata like Parts-Of-Speech (POS) tagging, Parse Tree, Named Entity Recognition (NER) etc. and stores them into MongoDB. We use NER for our framework.

Named Entity Recognition (NER) [16] locates and classifies named entities in text into predefined categories such as person, organization, location, etc. In our framework, we use person and organization to detect new actors. We use NER given by Stanford CoreNLP because it has higher accuracy [17], [18] than other similar tools like Alchemy [19], OpenCalais [20], OpenNLP [21], etc.

PropBank-based event coding. PropBank generates events [22] with 'predicate-argument' structures for each of the sentences of a document. In general, it has the arguments corresponding to the semantic roles of an agent/source (ARG0) and a target (ARG1, ARG2, etc.). It has extra arguments (annotation of modifier) that describe time (AM-TMP), location (AM-LOC), etc.

Table I shows the generated events for the following example.

"Obama vowed again on Sunday to help France hunt down the perpetrators of the attacks." - The Washington Post 11/15/2015

TABLE I: PropBank annotation example

Verb	ARG0	ARG1	AM-TMP
vowed	Obama	to help France hunt down the perpetrators of the attacks	again, on Sunday
help	Obama	France hunt down the perpetrators of the attacks	
hunt	France	down the perpetrators of the attacks	

The advantage of PropBank over Petrarch is in processing compound sentences. For example, consider the following sentence,

AFP reporter said Bashar-Al-Asad agreed to take help from President Trump and his Government.

Petrarch will give an event involving AFP reporter and Bashar-Al-Asad as source and target with the action "said". This is completely different from the actual event of interaction between Bashar-Al-Asad and President Trump. Propbank breaks this sentence into simpler form and it can capture the interaction between those two entities.

C. Political Actor Extraction

PropBank provides events with arguments. Each of them consists of a verb/action which is looked up from CAMEO dictionary to categorize a political event. Each political event contains arguments like ARG0 as a source, ARG1, and ARG2 as a target. We parse these arguments and filter name entities (Person, Organization) using NER that are missing in CAMEO actor dictionary. These actors are possible new actors. We detect an actor with alias names and group them together.

D. New Political Actor Recommendation

We introduce a window based actor ranking technique where we select top N actors for each window. We recommend an actor as a new actor if he/she appears in most of the windows. New actors role can be inferred from their co-occurred existing actors' roles. We describe them in details in section III.

E. Existing Actor's Role Change Recommendation

We integrate external knowledge bases (e.g., Wikipedia, Google Knowledge Graph [23]) into our framework to capture existing actor's role change. If any existing actor changes his/her role, then our framework captures the timeline change and recommend new roles if he/she has been assigned to new role (detail in section III-E).

F. Human Feedback

We provide a graphical user interface and a dashboard to the end user/human expert. It recommends new political actors with their role periodically. A user can easily view them and update actor dictionary (e.g., CAMEO) accordingly.

III. Recommending A New Actor and Role

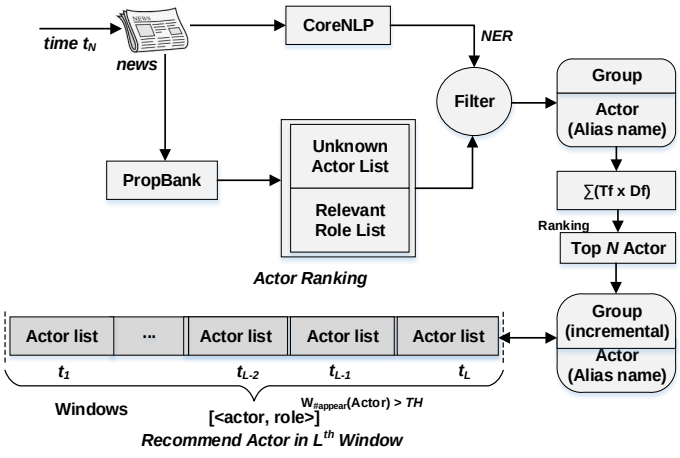


Fig. 2: Actor recommendation procedure in *RePAIR*

Figure 2 describes the technical details of our real-time actor recommendation technique. We use NER, PropBank to filter out all possibles actors from a news article. After that, we calculate a score or ranking for each actor across all the documents and select the top N actors. We infer their roles from their interacted (co-occurred in news article) existing actors' roles in CAMEO dictionary and suggest top M frequent roles. Here we have maintained a buffered time window W of length L. Each window contains the top N actors, which are merged with the previous windows actors list and updated the rank statistics. After L windows, we recommend new actors with their roles, if their occurrences in the L_{th} buffered window exceed a certain threshold TH. We describe this in details in forthcoming subsections starting with the following example.

A. New Actor Discovery

We use a frequency-based ranking algorithm to recommend the top political actors and their roles. Each news article/document contains raw text related to a political event. CoreNLP parses this raw text and outputs NER for each sentence. PropBank generates event code with arguments for each sentence using practnlp tool [24]. Each event code has an action/relation and arguments (ARG0, ARG1, ARG2, etc.) representing source/target of the relation. From these arguments, we filter out name entity list using NER. This list contains both existing actors and probable new actors. The role of an existing actor can be looked up from CAMEO dictionary. For example, *President Obama* has role $USAGOV$. NER gives us all actors (i.e Person, Organization) in a sentence. A news article contains multiple sentences. We filter out potential new actors by removing existing actors from NER actor list. Sometimes, an actor may come in different name/partial name across the article. For example, *Barack Hussein Obama* may come as *Obama* or *Barack Obama* in many sentences. As a result, we need to group these actors with all possible alias names and make a single actor name (e.g., *Barack Hussein Obama*). We try two similarity measures Levenshtein Distance (Edit distance) [25] and MinHash [26] independently to group actor's alias name. These similarity measures need a similarity threshold sim_{th} which is estimated from existing CAMEO dictionary (detail in section IV-B). After grouping, we calculate the term-frequency of each actor inside the news article. For an alias actor, we take the maximum count of the common phrase of an actor's name as its term-frequency because we group actor alias names. We assume that a common phrase from all alias names will be found throughout the document. For example, the term-frequency of *Barack Hussein Obama* will be the maximum frequency of *Barack Hussein Obama, Obama, and Barack Obama* as Obama is the common word between all names. Similarly, we group aliases for actors across multiple documents. Finally, for each actor a, we calculate a document frequency df and use the following equations to generate a score:

TABLE II: Symbol for algorithm

D	document set	d	document
m	meta-data	sim_{th}	similarity threshold
E	event code set	e	event code
A	actor set	a	actor
R	role set	r	role
tf	term freuency	df	document frequency
N	top N actor	L	number of window
k	merged actor as key	TH	Threshold in window

$$rank(a) = \sum_{d \in D} tf(a,d) \times df(a,D) \quad (1)$$

We use term frequency $tf(a,d) = count_{a \in d}(a)$ to show how frequent an actor a in a document d. Equation 1 shows the rank calculation of an actor a, where $df(a,D) = \frac{|d \in D : a \in d|}{|D|}$ is document frequency and it shows how frequent an actor a comes across all news articles/document in the set D.

B. Actor Role Discovery

Our framework extracts political events for sentences in a news article. We collect all events which contain new and existing political actors. CAMEO actor dictionary contains existing actors with their roles. Here, we assume that new actors' political roles will be related to their most frequent co-occurred existing political actors. So, we collect the roles and assign them to new actors as possible roles. We keep a map of each role with its occurrence. We update actors' maps with a role for all news articles. When an actor appears in two news articles, we include all roles in both articles and increase common roles' occurrence. While we have all ranked actors, we also have their role maps.

Algorithm 1 shows the new actor discovery algorithm with ranking. It takes the news document set D as input and gives actor map with rank M_{rank} as output. M_{rank} contains each actor as key with its ranking and role map M_{role}. It takes each document $d \in D$ and extracts new actor with roles R_d (lines 4-8). After that, it groups all alias names for each of the actors inside a document with counting term-frequency tf (lines 10-16). Next, it groups actors' alias names across multiple documents and updates the temporal actor map M which contains an actor as a key and a list of the tuple $< tf, d, R_d >$ as a value (lines 17-24). After that, for each actor in M, it calculates the document frequency d, and $rank$ score using equation 1 (lines 28-29). It then creates a role map M_{role} for each actor which contains all the possible roles with their total occurrences across the document set D (lines 31-35). Finally, we insert the actor as key and a tuple of $rank$, and role map M_{role} as value into M_{rank}. So, M_{rank} contains all possible new actors with their $rank$, and possible roles.

C. Recommending New Actors in Real-time

Our framework gets political news articles periodically and possesses a buffered time window W of length L. At each time window, it scans all the news articles and gets the possible actor list with rankings and possible roles. It is well known that new political actors will always be highlighted in media and people will always talk about them in print or online media. Here we assume that, if an actor comes in the top N ranking in multiple time windows, he or she has a high probability of being a new political actor.

Algorithm 1 New actor discovery

```
1:  procedure ACTORDISCOVERY(DocumentSet D)
2:      Actor Map M ← {}
3:      for each d ∈ D do
4:          m ← CoreNLP(d)
5:          E_d, A_current ← PropBank(d)
6:          A_all ← m.NER()
7:          A_d ← A_all − A_current
8:          R_d ← E.Roles()
9:          for each a ∈ A_d do
10:             tf ← count_{a∈d}(a)                    ▷ calculate tf
11:             for each a_i ∈ A_d − a do
12:                 if Match(a, a_i) > sim_th then
13:                     a ← Merge(A_d, a, a_i)
14:                     tf ← max(tf, count_{a_i∈d}(a_i))
15:                 end if
16:             end for
17:             k ← arg max_k{Match(M(k), a) > sim_th}
18:             if k ≠ empty then
19:                 k ← Merge(M, k, a_i)
20:                 M.insert(k, [(tf, d, R_d)])
21:             else
22:                 M.insert(a, [(tf, d, R_d)])
23:             end if
24:         end for
25:     end for
26:     M_rank ← {}
27:     for each k ∈ M.keys() do
28:         df ← |k ∈ d : d ∈ D|/|D|                   ▷ calculate df
29:         rank ← Σ_{tf∈M(k)}(tf) × df
30:         M_role ← {}
31:         for each R_d ∈ M(k) do
32:             for each r ∈ R_d do
33:                 M_role(r) ← M_role(r) + 1
34:             end for
35:         end for
36:         M_rank.insert(k, (rank, M_role)
37:     end for
38: end procedure
```

Algorithm 2 describes in details about the recommendation procedure. For each time window, we receive the actor list with ranking M_{rank}. We take the top N actors from the list (lines 5-7). We update a new actor Map M_A and its role map M_R. These are incrementally updated during each time window. For each actor in M_{rank}, we find the list of co-occurred actors in M_A. If there is no actor in M_A, we insert it in M_A with occurrence 1 and its role in M_R. If we find the closest match, then we merge the actor name if required. Later, we increment its occurrence in M_A. In the same way, we update the role of that actor in M_R (lines 8-20). As we increase the actors

Algorithm 2 Real-time new actor recommendation

1: **procedure** ACTORINREALTIME(TH_{min}, TH_{max}, N, L)
2: $New\ Actor\ Map\ M_A \leftarrow \{\}$
3: $New\ Actor\ Role\ Map\ M_R \leftarrow \{\}$
4: **for** $t \leftarrow t_1\ to\ t_L$ **do**
5: $D \leftarrow getDocuments(t)$
6: $M_{rank} \leftarrow ACTORDISCOVERY(D)$
7: $M_{topN} \leftarrow M_{rank}.top(N)$
8: **for each** $< a, (rank, M_{role}) > \in M_{topN}$ **do**
9: $k \leftarrow \arg\max_k \{Match(M_A(k), a) > sim_{th}\}$
10: **if** $k \neq empty$ **then**
11: $k \leftarrow Merge(M, k, a)$
12: $M_A(k) \leftarrow M_A(k) + 1$
13: **else**
14: $M_A(a) \leftarrow 1$
15: **end if**
16: **for each** $r \in M_{role}$ **do**
17: $M_R(k) \leftarrow M_R(k) + 1$
18: **end for**
19: **end for**
20: **end for**
21: **for each** $a \in M_A$ **do**
22: **if** $M_A(a) \geq TH_{max}$ **then**
23: $M_A(a) \leftarrow$ "$new\ actor$"
24: $M_R(a) \leftarrow$ "$new\ actor's\ role$"
25: **end if**
26: **if** $M_A(a) \leq TH_{min}$ **then**
27: $M_A(a) \leftarrow$ "$discard\ actor$"
28: $M_A.remove(a)$
29: $M_R.remove(a)$
30: **end if**
31: **end for**
32: **end procedure**

in M_A and M_R, their size will always be increased. So, we introduce a (min, max) threshold $TH_{min}\ and\ TH_{min}$. After the L^{th} time window, if the total occurrence of an actor is greater that TH_{max}, we consider his/her as a new actor and recommend top N roles from M_R also. On the other hand, it total occurrence is below TH_{min}, we discard the actor from M_A and M_R (lines 21-31). After L^{th} time window, we will recommend new actors with possible related roles. Human experts will verify this and update CAMEO dictionary accordingly.

D. Graph-based Role Detection Technique

We introduce a graph-based technique to suggest roles for recommended actors. This is an alternative to our frequency based role detection technique. Roles of a new actor will be influenced by existing actors with whom he/she has mostly interacted. Therefore, we use weighted label propagation technique[27] to infer possible roles from existing related political actors.

We formulate a graph $G = (V, E)$ that implies the interaction between actors. Here, V is the set of actors (contains both existing and recommended actors). For two actors, u and v, $(u, v) \in E$ represents those actors are mentioned in the same news article. We also assign a weight function $w(u, v)$ that implies how frequent u and v are co-occurred in the news articles.

$$w(u, v) = \text{co-occurrence count of } u \text{ and } v$$

After that, we assign labels to the nodes which are the roles found in the CAMEO dictionary for existing actors. For recommended actors, we simply put an empty label. We use $label(x)$ to denote the label for actor x.

$$label(x) = \begin{cases} \text{roles from dictionary} & \text{if } x \text{ is an existing actor} \\ empty & \text{otherwise} \end{cases}$$

Now we run our iterative label propagation algorithm, where we assign weight to each possible role for a recommended user. Roles are selected from neighboring actors. Weight is assigned for each role based on a weight associated with the edge of interaction graph, G. We repeat the process N times or until the label assignments become stable. The weighting function for roles works as follows,

$$role - weight(u, role) = \sum_{v \in Neighbors(u)} T(v, role) \times w(u, v)$$

where $T(v, r)$ is used as indicator variable that simply returns 1 if v has 'r' in it's list of roles, 0 otherwise. In figure

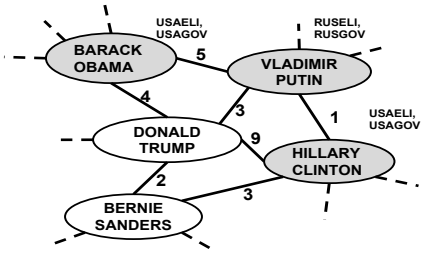


Fig. 3: Example Scenario for Graph Based Role Recommendation

3, we will determine the role for DONALD TRUMP. He is associated with both existing actors (gray nodes) and new political actors. All the existing actors have their roles associated with their nodes. For example, BARACK OBAMA has USAELI and USAGOV as his roles. Now, we use our role-weight method to calculate the role. For example,

$$\text{role-weight(DONALD TRUMP, USAGOV)}$$
$$= 4 \times 1 \ + \ 9 \times 1 \ + \ 3 \times 0 \ + \ 2 \times 0 \ = \ 12$$

After calculating weights for all roles from neighbors, we find USAGOV, USAELI, RUSGOV as the top 3 possible roles for DONALD TRUMP.

E. Integrating External Knowledge Bases

External knowledge bases (e.g., Wikipedia, Google Knowledge Graph, etc.) store up-to-date news and they can be helpful for recommendation. Although they are the big resources but they are not alternatives to CAMEO dictionary

because CAMEO dictionary contains more structured data (e.g., action/actor dictionary related to politics) especially for political event coding where knowledge bases contain generic and unstructured data. Moreover, as both of them are curated by humans, external knowledge bases suffer the same problem as CAMEO if any new political actor appears. So, we extend our framework to integrate these external knowledge bases to identify any timeline change in role for the existing actors. For each actor in the CAMEO dictionary, we query Google Knowledge Graph [23] for the appropriate entity. It helps us to disambiguate between closely matched entities while searching via popular words (i.e name of politicians, media personals, etc.). The actor name from dictionary may fail to match with Wikipedia database in many cases but Knowledge Graph helps us to find the appropriate search text. For example, if we search for *Mullah Mohammad Rabbani* using Wikipedia API, we don't find any page related to him. Knowledge Graph resolves him to *Mohammad Rabbani* and then we are able to locate him in the Wikipedia database. We observe a samples of 1000 actors from a total of 18000 actors and find only 20 of them are resolved only with Wikipedia whereas 850 are correctly identified when we use Knowledge Graph. We identify whether there is a change in the timespan for the actor by observing the dates associated with each of his/her responsibilities based on the information inside Infobox [28] of the Wikipedia page. We list the actors for whom we detect a change in the timeline of the role. For example, former President of United States, Barack Obama has a dictionary entry where his role is USAGOV and started on 20 January 2009 but has no end-date. From Wikipedia InfoBox, we find both start date and end date of his presidency. We match start date with the existing entry of the dictionary and add the end date to the list of updates. We also suggest a new role for an existing actor if he/she is assigned to a new political position based on the information inside the Wikipedia Infobox. Wikipedia gives us details about the roles (e.g., President of United States, Prime Minister of India, etc.) for each existing political actors and we get the short form of code for the corresponding actor's roles in CAMEO dictionary (e.g., USAGOV, USAMIL, etc.).

IV. Experiments

A. Setup and Dataset

To evaluate our framework, we take 10 time windows at 24 hours interval. In each time window, we scrap newly published online news articles listed in 400 RSS feeds around the world. We process these articles/documents with CoreNLP, practnlptools running in Apache Spark. Finally, the processed articles are used for recommendation related tasks. We have in-total of 131,932 news articles for the experiment.

B. Threshold Estimation

We are using a similarity measure for grouping all actors with their alias names. This uses a partial string matching. In the experiments, we use MinHash and Edit distance and show their performance independently. Both approaches give a numeric value between 0 and 1 indicating the similarity. For our purpose, we need to set a threshold value for considering whether two strings represent the same actor using these methods. When the similarity value between two actor names becomes greater than the threshold, we regard them as the same entity. To estimate the threshold, we consider the similarity values for actors already in the CAMEO actor dictionary. As an example, consider the following snippet from the actor dictionary for Kofi Annan and Boutros Boutros Ghali:

```
....
KOFI_ANNAN_
+SECRETARY-GENERAL_KOFI_ANNAN
.....
BOUTROS_BOUTROS_GHALI_
+BOUTROS_BOUTROS-GHALI_
....
```

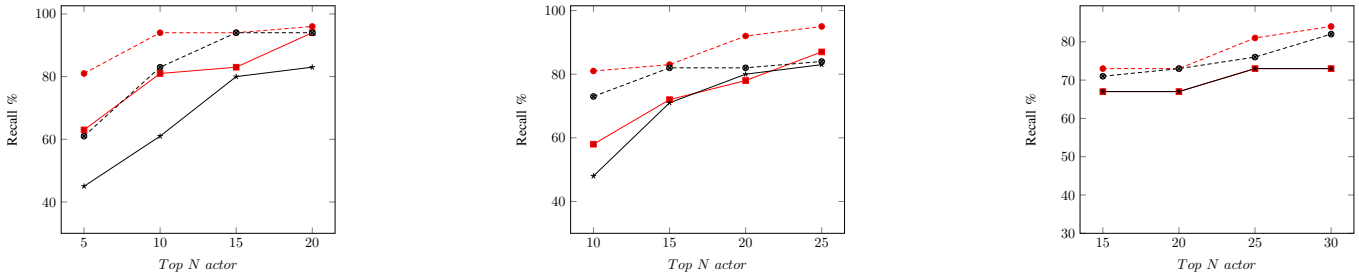
Both of them are followed by multiple aliases (lines starting with a '+'). We calculate the similarity between these aliases and the aliases of different actors. In the first case, we observe higher values (close to 1). For the later case, we found smaller values. Then we set a value that will minimize the number of false positives (i.e., reporting high similarity when they are not same actor). By this empirical study, we estimate thresholds for Edit distance and min-hash based methods to be 0.75 and 0.4 respectively.

C. Baseline Methods

We consider the following methods as baselines.

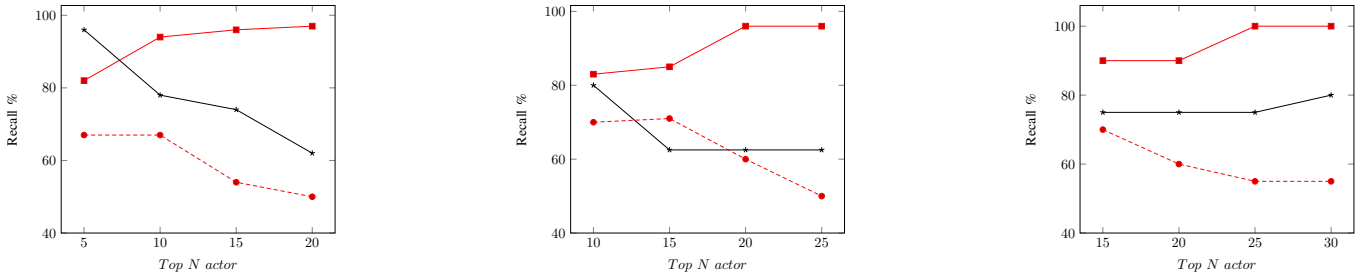
PETRARCH [2] takes fully-parsed text summaries in Penn Tree format [8] from Stanford CoreNLP and generates events in a 'who-did-what-to-whom' format. It outputs source, target and a CAMEO code [29] for the political event along with other information like the date of an event, news source, etc. We used it as a baseline event coder and apply a frequency based approach to recommend new actors. We use it to show how semantic role labeling (used in our framework) improved performance. We compare it with our approach in terms of both precision and recall.

BBN ACCENT$^{\mathbf{TM}}$ [30] event coder automatically extracts event data from news reports from around the world. It is based on $BBN\ SERIF^{TM}$, a natural language analysis engine that extracts structured information (e.g. entities, relationships, and events) from a text. Its events are coded according to the Conflict and Mediation Event Observations (CAMEO) ontology. It is a proprietary political event coding tool developed by Raytheon BBN Technologies Corp [31]. BBN encodes political event which has actor, agent, and action (CAMEO event type) but it does not generate the role of an actor. As BBN is proprietary and its used dictionaries are encrypted, so our comparison is based on new actor recommendation with human expert validation, i.e., precision comparison. This means that our framework will recommend new actors using Propbank, PETRARCH, and BBN separately and a human expert will validate how many of them are really new actors for each of the event coding tools.



(a) Deleted actor = 5 (b) Deleted actor = 10 (c) Deleted actor = 15

Fig. 4: Performance for Actor recommendation. Recall: Edit distance (PropBank) --●--, MinHash (PropBank) --⊗--, Edit distance (PETRARCH) —■—, MinHash (PETRARCH) —⋆—



(a) Deleted actor = 5 (b) Deleted actor = 10 (c) Deleted actor = 15

Fig. 5: Performance for role recommendation. Recall: Edit distance —■—, MinHash —⋆—, Exact match --●--

D. Experiment 1: Performance Evaluation

This experiment evaluates the framework's correctness. To do this, we first select 30 existing political actors from the dictionary and randomly remove some of them. We start with removing 5, 10 and 15 actors. The deleted actors are considered for a recommendation (as if they are newly discovered). We calculate how many of them are recommended by our algorithm. This gives us the recall. We repeat the experiments 10 times and get averages for 3 different numbers of deleted actors and plot the result in Figure 4 We can not compute precision because all the recommended actors are political actors. The retrieved actors are 15-20% of the recommended actors.

From Figure 4, we see that if we increase the top N actors (i.e., how many actors are recommended at each time window), it will increase the possibility of retrieval of a deleted actor. Because there are actors who are not highly focused by media but have a continuous impact in his/her locality.

We also see that our framework which has PropBank based event coder (dotted line) has higher recall than PETRARCH based event coder (solid line) for recommending new actors. This is expected because PETRARCH misses many verbs as event while coding. For example, Table I in section II-B shows three separate events using PropBank but PETRARCH will generate only one event 'FRA(Source), USAGOV(Target), Cooperate(Verb)'. As a result, our framework will generate more new actors that PETRARCH. Moreover, Edit distance based (red) actor alias grouping has higher recall than MinHash based (black) actor alias grouping for both approaches because Edit distance considers each character where MinHash consider each word while comparing alias names of an actor.

In the following experiment, we suggest the possible roles for identified actors in the previous experiment. We follow the similar sampling process as the experiment above. As a reference, we have their roles (R1) listed in the CAMEO dictionary. Now using our frequency-based role recommendation algorithm we predict appropriate roles (R2) for them. We calculate the intersection of the two sets, R1 and R2. If the output is non-empty set then we consider it as a success. For the purpose of the experiment, we keep the size of R2 as 5. The roles are selected based on frequency. Top 5 frequent roles are presented as possible actor roles. Again we vary the number of deleted actors. We also vary the top-N. For the experiment, we reported the ratio of a number of success and number of actors retrieved. Figure 5 shows the results. We see that Edit distance has a slight advantage while actor recommendation. We use word set in MinHash calculation and it uses Jaccard similarity. So, for close matched actor aliases like 'Donald Trump' and 'Donald J Trump', Edit distance gives higher similarity than MinHash. For similar reason, we see recall variation in figure 5 for role recommendation. Roles of an actor is a short text. Edit distance performs better when top N increases, whereas MinHash and Exact matching have similar level of performance (worse than Edit distance). Because for Edit distance, we consider partial matching. For example, USAMILGOV and USAMIL are similar using Edit

distance method, but hugely different when MinHash (Jaccard similarity) is used. They are certainly different for Exact Matching.

Discovering roles with Word2vec. We use Word2vec to find the related roles, which requires a training model. We take possible new actors with related existing actors from news articles. We collect them across 1 *to* L^{th} (algorithm 2) time windows to train the model. For each new actor, we find the top N similar existing actors (in the CAMEO dictionary) using Word2vec's $similal\_word$ method and recommend their roles. Unfortunately, we get less than 5% recall. This is obvious because Word2vec does not apply alias actor grouping. As a result, it fetches less similar actors than our frequency-based approach.

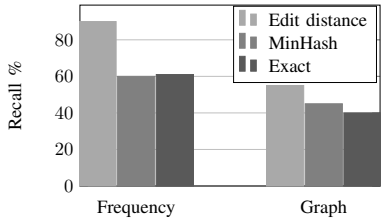


Fig. 6: Comparison of actor role recommendation with baseline: (N = 15, deleted actors = 15)

Comparison of role recommendation techniques. In this experiment, we first delete some well-known actors from the existing CAMEO dictionary. Then we try to retrieve their roles by frequency based and **graph based** role detection techniques. We fix the values of numbers of recommended actors per window and number of deleted actors to 15. We do not vary those parameters for graph-based approach because it caused the resulting graph to be very sparse. That in-turn makes the role inference hard with label propagation technique. We consider all the actors suggested by the actor recommendation algorithm and their interactions to formulate the graph. Figure 6 shows the statistics for the experiment. Here, frequency based approach outperforms graph-based approach for all the similarity measurement techniques because it infers new actor's role from existing actors' roles. But in the case of graph-based approach, it considers roles from neighbors who can be either existing or new actors. In that case, the error with one role assignment can propagate to others. Label propagation algorithm have less number of labels to propagate and the edges between new actors will be higher because they have a larger possibility to co-occur in a document.

E. Experiment 2: Recommendation of New Actors with Roles

Here, we list the possible newly recommended political actors based on a threshold. The threshold is how many times an actor appeared in the time windows. We set the number of time window L = 10 and (max, min) frequency threshold (TH_{max}, TH_{min}) = (5, 2) in algorithm 2. So if any actor appears in 5 (more than 50% window length), he/she will be suggested as a potential new political actor. In the case of roles, we list the most probable roles from their co-occurred existing political actors. Table III shows the list of recommended actors after all windows. We use both MinHash and Edit distance based string similarity and list the output side-by-side. We found that both the approach detects similar roles for identical actors. But the recommended user list varies.

Compare with PETRARCH and BBN ACCENT. We use PETRARCH and BBN ACCENT as event coder in our framework and run experiments to recommend new actors. Human expert validates recommended new actors and we calculate precision based on that. We run the same experiment using our framework and calculate precision. We vary the number of top N recommended actors. Figure 7 shows the comparison. It shows that our framework (PropBank based event coder) has higher precision (dotted red) than PETRARCH (solid red) and BBN ACCENT (solid black) coding. If BBN ACCENT fails to find any actor in the dictionary, it generates an empty event. On the other hand, PETRARCH separates political and nonpolitical events with actors but misses many verbs as event compares to PropBank based coding.

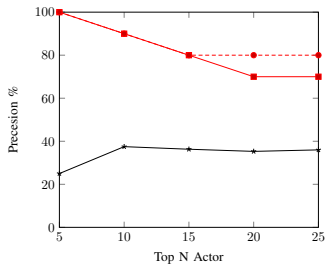


Fig. 7: Baseline coding comparison in actor detection: PETRARCH —■—, BBN ACCENT —⋆—, and PropBank --●--

F. Experiment 3: Scalability Test

We ran our experiment on a Spark cluster which has 1 master node and 10 slave nodes. Each node has 8 vCPU, 16 GB memory and 1 TB space.

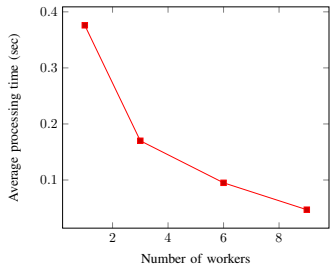


Fig. 8: Processing time for burst input of 131,932 documents

We are using CoreNLP which is computationally expensive [15]. Each CoreNLP instance runs inside a Spark worker node and extracts metadata from the news article. Figure 8 shows the scalability of average document processing time of 131,932 news articles using Spark. Spark fetches the news articles at 10 minutes intervals, whereas the scrapper sends news at 24 hours interval. For this test, we take 131,932 news articles in a burst to observe actual performance. Normally, the scraper fetches 5000-6000 online news article at the beginning of a

TABLE III: List of recommended actors with their roles

Edit distance		MinHash	
Actor	**Top 3 roles**	**Actor**	**Top 3 roles**
DONALD TRUMP	USA, USAGOV, GOV	DONALD TRUMP	USA, USAGOV, GOV
EMMANUEL MACRON	USAGOV, FRA, FRAGOV	EMMANUEL MACRON	USAGOV, FRA, FRAGOV
MITCH MCCONNELL	GOV, USA, USAGOVLEG	REX TILLERSON	USAGOV, USA, QAT
REX TILLERSON	USAGOV, USA, QAT	MITCH MCCONNELL	GOV, USA, USAGOVLEG
LIU XIAOBO	MED, CHNOPP, CHN	RODRIGO DUTERTE	PHLGOV, LEG, GOV
RODRIGO DUTERTE	PHLGOV, LEG, GOV	YEMI OSINBAJO	NGAGOV, NGA, GOV

run. Later, it scrapes only the updated news articles which are fewer than initial. Spark internally creates micro-batches to process these 131,932 articles. So, we take average processing time per article. If we have only one worker, then its average processing time is almost 0.376 sec which is reduced to 0.047 when the total number of worker reaches to 10 (maximum load). It scales up almost linearly.

V. Related work

Political event data analysis has been developed over multiple decades for international relations and security studies. Saraf et al. [32] showed a recommendation model to determine if an article reports a civil unrest event. Schrodt and Van Brackle [33] show a complete picture of generating events from raw news texts. While each of these works focuses on political event data generation and analysis, none incorporate dynamic actor dictionary building. Here, we design a real-time scalable framework that detects new actors dynamically.

Role recommendation based on surrounding keywords in the document can be considered. For example, 'President' keyword occurring in front of Barack Obama may help infer his role. But we may not get enough keywords like above to infer roles. Also often they do not convey the complete role for an actor. Our above example shows similar scenario. With the keyword 'President', we can imply Barack Obama as a government employee. But we cannot infer his country. Moreover, CAMEO uses short-form encoded roles (e.g., USAGOV) which require a static mapping if we consider the above procedure. [34], [35], [36] discuss a procedure to discover roles in a large network. It completely discovers roles for every node in the graph. In our case, nodes representing existing actors have already assigned a role. Also considering structural similarity to role prediction will not be a feasible idea. Because actors with similar roles do not necessarily involves in similar structure in graph.

Automatic Content Extraction (ACE) based event extraction systems use pattern or machine learning. Ritter et al. [37] proposed TWICAL - an open domain event extraction and categorization system for Twitter. Georgescu et al. [38] showed event extraction using Wikipedia. Weninger et al. [39] shows future Web content extraction algorithms. All the above methods use classifiers to identify event type which performs well in a generic domain but not in a restricted political domain. Therefore, we use a mixed model of CAMEO and PropBank to extract political events.

Extracting events from raw news articles in real-time demand a scalable distributed streaming framework. We choose Spark Streaming because it is matured and widely used in industry.

VI. Conclusion and future work

In this paper, we address the problem of detecting and recommending new political actors and their roles in real-time. We propose a Spark-based framework called *RePAIR* with unsupervised ranking techniques with new actor aliases grouping that works on news articles collected on a periodic basis to recommend new actors. Moreover, we integrate external knowledge bases (Wikipedia, Infobox, etc.) to capture the timeline change for existing actors and also suggest roles for them if they have new political roles. Our experimental results evaluate the framework's performance.

Currently, we limit ourselves to find new political actors but this approach can be extended to recommend new political actions in CAMEO verb dictionary. In addition, we will extend this to build CAMEO dictionaries for other languages (e.g., Spanish, Arabic).

References

[1] J. Beieler, P. T. Brandt, A. Halterman, P. A. Schrodt, and E. M. Simpson, "Generating political event data in near real time: Opportunities and challenges," *Computational Social Science: Discovery and Prediction. ed. by R. Michael Alvarez, Cambridge, Cambridge University Press*, pp. 98–120, 2016.

[2] *PETRARCH*, http://petrarch.readthedocs.org/en/latest/.

[3] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation."

[4] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[5] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *17th International Conference on ACL, Volume 1*, 1998, pp. 86–90.

[6] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[7] E. Joanis and S. Stevenson, "A general feature space for automatic verb classification," in *10th conference on European chapter of ACL, Volume 1*, 2003, pp. 163–170.

[8] *The Penn Treebank Project*, The University of Pennsylvania, https://www.cis.upenn.edu/~treebank/.

[9] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," *9th USENIX*, pp. 2–2, 2012.

[10] *Stanford CoreNLP*, Stanford, http://nlp.stanford.edu/software/corenlp.shtml.

[11] *MongoDB*, MongoDB, https://www.mongodb.com.

[12] J. Kreps, N. Narkhede, J. Rao *et al.*, "Kafka: A distributed messaging system for log processing." NetDB (pp. 1-7), 2011.

[13] *Web Scraper*, Open Event Data Alliance, http://oeda-scraper.readthedocs.io/en/latest.

[14] *RSS Whitelist*, https://github.com/openeventdata/scraper/blob/master/whitelist_urls.csv.

[15] M. Solaimani, R. Gopalan, L. Khan, P. T. Brandt, and B. Thuraisingham, "Spark-based political event coding," in *BigDataService*. IEEE, 2016, pp. 14–23.

[16] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[17] K. J. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska, "Comparison of named entity recognition tools for raw ocr text." 2012.

[18] S. Atdağ and V. Labatut, "A comparison of named entity recognition tools applied to biographical texts," in *ICSCS*. IEEE, 2013, pp. 228–233.

[19] *AlchemyAPI*, IBM Watson, http://www.alchemyapi.com/.

[20] *Open Calais*, Thomson Reuters, http://www.opencalais.com/.

[21] *Apache OpenNLP*, The Apache Software Foundation, http://opennlp.apache.org/.

[22] O. Babko-Malaya, "Propbank annotation guidelines," 2005.

[23] *Google Knowledge Graph*, https://developers.google.com/knowledge-graph/.

[24] *practnlptools 1.0*, https://pypi.python.org/pypi/practnlptools/1.0.

[25] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[26] A. Z. Broder, "On the resemblance and containment of documents," in *Compression and Complexity of Sequences 1997. Proceedings*. IEEE, 1997, pp. 21–29.

[27] H. Lou, S. Li, and Y. Zhao, "Detecting community structure using label propagation with weighted coherent neighborhood propinquity," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 14, pp. 3095–3105, 2013.

[28] *Wikipedia Infobox*, https://en.wikipedia.org/wiki/Help:Infobox.

[29] *Conflict and Mediation Event Observations (CAMEO) Codebook.*, CAMEO, http://eventdata.parusanalytics.com/data.dir/cameo.html.

[30] E. Boschee, P. Natarajan, and R. Weischedel, "Automatic extraction of events from open source text for predictive forecasting," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 51–67.

[31] *Raytheon BBN Technologies.*, http://www.raytheon.com/ourcompany/bbn/.

[32] P. Saraf and N. Ramakrishnan, "EMBERS autogsr: Automated coding of civil unrest events," in *22nd ACM SIGKDD*, 2016, pp. 599–608.

[33] P. A. Schrodt and D. Van Brackle, "Automated coding of political event data," in *Handbook of Computational Approaches to Counterterrorism*. Springer, 2013, pp. 23–49.

[34] R. A. Rossi and N. K. Ahmed, "Role discovery in networks," *CoRR*, vol. abs/1405.7134, 2014. [Online]. Available: http://arxiv.org/abs/1405.7134

[35] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: structural role extraction & mining in large graphs," in *18th ACM SIGKDD*, 2012, pp. 1231–1239.

[36] R. A. Rossi, B. Gallagher, J. Neville, and K. Henderson, "Modeling dynamic behavior in large evolving graphs," in *6th ACM WSDM*, 2013, pp. 667–676.

[37] A. Ritter, O. Etzioni, S. Clark *et al.*, "Open domain event extraction from twitter," in *18th ACM SIGKDD*, 2012, pp. 1104–1112.

[38] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer, "Extracting event-related information from article updates in wikipedia," in *European Conference on Information Retrieval*. Springer, 2013, pp. 254–266.

[39] T. Weninger, R. Palacios, V. Crescenzi, T. Gottron, and P. Merialdo, "Web content extraction: a metaanalysis of its past and thoughts on its future," *ACM SIGKDD Explorations Newsletter*, vol. 17, no. 2, pp. 17–23, 2016.