CoMe-KE: A New Transformers Based Approach for Knowledge Extraction in Conflict and Mediation Domain

Erick Skorupa Parolin*, Yibo Hu*, Latifur Khan*, Javier Osorio‡, Patrick T. Brandt†, Vito D'Orazio†
Department of Computer Science, School of Economic, Political, and Policy Sciences†*
The University of Texas at Dallas, Richardson, Texas
School of Government and Public Policy‡, *University of Arizona*, Tucson, Arizona
{erick.skorupaparolin, yibo.hu, lkhan, pbrandt, dorazio}@utdallas.edu, josorio1@arizona.edu

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9672080

***Abstract*—Knowledge discovery and extraction approaches attract special attention across industries and areas moving toward the 5V Era. In the political and social sciences, scholars and governments dedicate considerable resources to develop intelligent systems for monitoring, analyzing and predicting conflicts and affairs involving political entities across the globe. Such systems rely on background knowledge from external knowledge bases, that conflict experts commonly maintain manually. The high costs and extensive human efforts associated with updating and extending these repositories often compromise their correctness of. Here we introduce CoMe-KE (Conflict and Mediation Knowledge Extractor) to extend automatically knowledge bases about conflict and mediation events. We explore state-of-the-art natural language models to discover new political entities, their roles and status from news. We propose a distant supervised method and propose an innovative zero-shot approach based on a *dynamic hypothesis* procedure. Our methods leverage pre-trained models through transfer learning techniques to obtain excellent results with no need for a labeled data. Finally, we demonstrate the superiority of our method through a comprehensive set of experiments involving two study cases in the social sciences domain. CoMe-KE significantly outperforms the existing baseline, with (on average) double of the performance retrieving new political entities.**

***Index Terms*—knowledge base construction, knowledge extraction, ontologies, link and graph mining, transfer-learning, natural language processing, web search and mining, semantic-based data mining, CAMEO**

I. Introduction

The recent advances in artificial intelligence (AI) and natural language processing (NLP) enable developing sophisticated semantic applications, opening up new perspectives for knowledge discovery and extraction from very large (un)structured databases. These search mechanisms, expert and diagnosis systems, and information extraction frameworks are employed successfully in many areas. Such technologies are supported by intelligent systems, frequently utilizing ontological background knowledge and knowledge bases/graphs.

In the political and social sciences, scholars and governments dedicate large significant resources analyzing the global interactions among political entities, designing forecasting models of conflict, and monitoring current affairs to predict trends. Several ontologies and knowledge bases for this purpose have been developed and used widely in political science, such as CAMEO (**C**onflict **a**nd **M**ediation **E**vent **O**ntology) [1], WEIS (World Event Interaction Survey), and COPDAB (Conflict and Peace Data Bank) [2]. Recent works in this domain typically focuses on event extraction based on classical machine learning [3]–[7], deep learning [8]–[11], sophisticated language understanding models [12]–[16] and rule-based approaches [17]–[19]. The majority of these efforts leverage ontology components and/or rely on the aforementioned knowledge bases to extract events from text.

Here we focus on the CAMEO since it currently is the most prominent event coding framework in the political science and international relations communities. It provides an ontology and extensive knowledge bases, facilitating the extraction of structured events. In practice, CAMEO incorporates extensive repositories mapping political entities to the corresponding roles they occupy in a region (e.g., country level), ethnicity, or religion.

Constructing, maintaining, and updating CAMEO-like knowledge bases and repositories have been an extremely costly and time-consuming process that requires exhaustive human efforts and specialist knowledge. As time passes and new political entities appear in the news, scholars face a considerable cost working with outdated knowledge bases and repositories. To address these obstacles, we seek alternative solutions from other semantic computing sub-areas such as knowledge base construction and open information extraction.

Automatic knowledge base construction (KBC) is an established task largely employed with information extracted from documents. Some previous works [20]–[28] focused on exploring encyclopedic knowledge utilizing pre-defined format of entities and relations, which usually male the application unfeasible when dealing with domain specific schemes, such as CAMEO. Others [29], [30] extend existing domain specific ontologies and knowledge bases/graphs, relying on their pre-existing entries to learn new entities and relations. These approaches are not applicable for extending ontology toward close or related domains though.

Open information extraction (OpenIE) is an alternative NLP task that generates structured representations of information

in triples or n-ary propositions from text. However, given the open-domain nature of this task, previous work [31]–[35] captures too wide a range of knowledge and relations, hindering the adaptation and application of these approaches on domain specific pre-defined ontologies. Further some of these approaches only capture the knowledge that is explicitly mentioned in text, ignoring contextual semantics.

For the political conflict and mediation domain, to the best of our knowledge, RePAIR [29] is the only existing framework particularly designed to extend automatically the CAMEO repositories. RePAIR has the limitations that may constrain its application in some cases. First, despite high recall retrieving new political entities, it fails on mapping the relations between these entities and their corresponding roles and statuses. Second, RePAIR does not allow extending CAMEO to other relevant sub-domains in political science that currently are not covered by this ontology (e.g., organized crime, voting behavior, judicial behavior, parliamentary systems, congress or legislative behavior, etc.). Finally, it does not allow extending or including new semantic relations to the ontology.

Recent advances in deep neural networks and language understanding models open new opportunities and address the challenges in our task. Pretrained language models (PLMs) based on Transformers [36] structures (such as GPT [37], BERT [38], and RoBERTa [39]) have demonstrated substantial gains in challenging NLP tasks by leveraging transfer learning techniques. This paradigm allows the pre-training of complex models in a large corpus and later fine-tune on specific downstream tasks, requiring only small (or even no) labeled data, and alleviating the human annotation bottleneck for specific applications. Therefore, PLMs are expected to boost the performance of distant-supervised methods [40]–[42] for relation extraction in weakly-labeled domain corpora. Moreover, PLMs greatly improve task-agnostic few-/zero-shot performance [43]–[47] by exploiting the domain knowledge already encoded within them through natural language inference formulation [48].

This paper introduces the CoMe-KE (**Co**nflict and **Me**diation **K**nowledge **E**xtractor) framework leveraging PLMs to address the challenges associated with CAMEO's knowledge base extension. CoMe-KE implements two modules based on distant supervised (DS) and zero-shot (ZS) learning to extract the roles and statuses of actors from text. We explore such learning techniques given the lack of labeled data, which is a common restriction in the political and social sciences.

This work includes the following contributions: First, we explore state-of-the-art natural language understanding models through transfer learning to design a two-step framework to extract CAMEO-like knowledge from text. Second, we design a DS-based approach with transformer models to detect the political roles of entities. Third, we propose an innovative ZS-based method to for extract relations via our dynamic hypothesis procedure (discussed in Subsection IV-B2). Our ZS-based method explores other social science sub-domains not covered in CAMEO and extracts new semantic relations expanding the ontology. Finally, we explore two comprehensive study cases in political science to demonstrate CoMe-KE's superiority.

The rest of the paper is structured as follows. Section II details the CAMEO ontology and provides background. Section III defines the problem addressed the current modeling strategy. Section IV provides a detailed design of the CoMe-KE, while Section V specifies the source, volume and other features about the datasets utilized in our experiments. Section VI shows the results, and Section VII concludes.

II. CAMEO Codebook

CAMEO is a dominant ontology for political event data, that incorporates a knowledge base (KB) of actors (or entities) dictionaries ($\approx$ 67K entries) and action-pattern dictionaries ($\approx$ 14K verb phrases). The former is a repository for political entities, such as domestic, international, military non-state, and general political actors / agents, while the latter store representations of their political actions or interactions.

The actor dictionaries are divided into two major types of entities: domestic and international. While domestic actors are linked to countries through domestic relationships associated to the role they play, the international entities are international governmental organizations, militarized groups, non-governmental organizations, non-governmental movements, or multinational corporations. Besides these relationships, such entities also contain attributes that may be relevant on conflict and mediation analysis, such as ethnicity, religion and party.

The official CAMEO codebook annotates 26 distinct hierarchical roles.[1] We analyze the most relevant (and frequent) roles for the government domestic entities:

- GOV: generic government role indicating participation in the executive, governing parties, etc;
- GOVBUS: entities about finance activities;
- GOVDEV: entities about development issues;
- GOVEDU: entities dealing with education;
- GOVELI: elites and former government officials;
- GOVENV: entities focusing on environmental and ecological issues;
- GOVHLH: entities about health and social welfare;
- GOVHRI: entities dealing with human rights concerns;
- GOVLAB: entities concerned with labor issues;
- GOVLEG: parliamentarians, lawmakers, senators, etc;
- GOVMED: entities for mass information dissemination;
- GOVREF: entities about migration and relocation.

For example, Senate Majority Leader [GOVLEG] of United States [USA] Charles Ellis Schumer currently should appear in dictionaries with the code USAGOVLEG; while Brazilian Minister of Education Milton Ribeiro should be coded BRAGOVEDU. Since CAMEO is not up to date, neither of these are current in the repositories.

Let the term **role-code**, φ_p, denote the (domestic) role (last characters in the code) that a given actor p performs in a country (first three characters of the code) s/he represents.

[1] https://parusanalytics.com/eventdata/data.dir/cameo.html

III. Problem Definition and Modeling Solution

Given a natural language input corpus, we want aim to extract automatically new political entities, associate them to their country, and identify a role they perform in their political environment. This collects tuples of (*actor*, *role-code*) that serve as entries for CAMEO's repositories. A model based on two tasks is proposed: (i) recognize the potential political entities and their locations from text, and (ii) identify the role for a potential political entity previously retrieved in task (i).

We formulate the first task as *named entity recognition* (NER), where named entity chunks are classified as *person* (PER) to consider potential political actors, and as *locations* (either GPE or LOC) for possible regions where they act. Thus, for each input document X, we combine the sets of m potential political entities and n locations extracted from X through a cross-product operation to obtain the collection $C_X = \{(p_1, l_1), (p_1, l_2), \ldots, (p_m, l_n)\}$ of size $|C_X| = m \times n$.

The second task is formulated as *relation extraction* (RE), assigning a semantic role to each *person-location* tuple in C_X. Formally, given a X and the collection C_X, we want to learn the function $f_\theta(X, p_i, l_j)$ that will map any tuple $(p_i, l_j) \in C_x$ to a categorical role $r \in P$, corresponding to the relation expressed in X between the entities p_i and l_j. Here P = {GOV, GOVBUS, GOVDEV, GOVEDU, GOVELI, GOVENV, GOVHLH, GOVHRI, GOVLAB, GOVLEG, GOVMED, GOVREF, NOTHING} is the set of roles described in Section II plus the relation *NOTHING* for those cases where the potential political entity p_i is unrelated to the country l_j (i.e., majority of the cases).

A major constraint in this design is the lack of labeled data, which is a recurrent restriction while working in political and social sciences. But the existing CAMEO entries may be useful for automatically learning how to generate new entries. Therefore, we explore two learning strategies in our proposed methods: (i) DS learning by leveraging the existing knowledge base, and (ii) ZS learning. We expand the design and components of the framework in Section IV.

IV. Proposed Framework

As described in Section III, we propose a modeling solution based on two NLP tasks (NER and RE) to extract knowledge from news articles and extend the CAMEO knowledge bases.

Figure 1 illustrates the framework design. CoMe-KE collects political news through web scrapers and applies certain pre-processing procedures. Later, it parses the text to extract entity chunks corresponding to person names (potential political actors), geopolitical and general locations through an NER model. Next, one of two strategies for relation extracting steps is chosen: the flow toward the top of Figure 1 illustrates the ZS approach, while the flow to the bottom shows the DS method to extract political entities and their roles in the countries. In the following subsections elaborate the CoMe-KE's components.

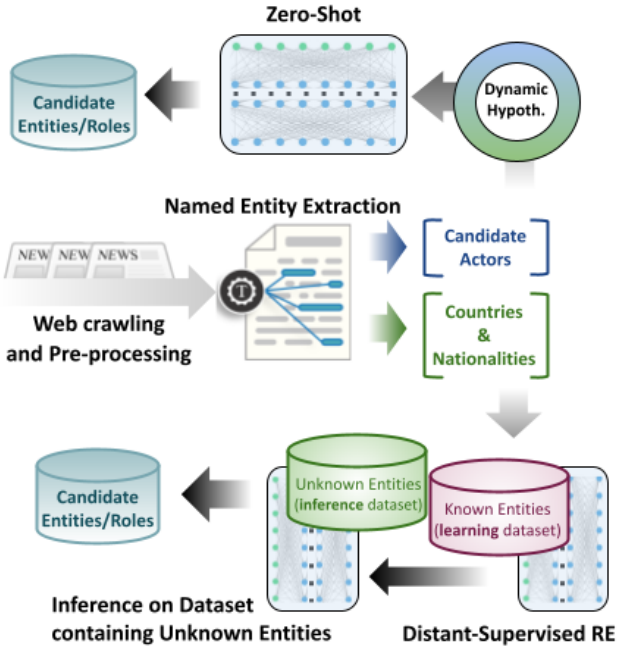


Fig. 1: CoMe-KE design: modular implementation to extract knowledge from news articles via DS or ZS learning.

A. Named Entity Recognition Component

From Algorithm 1, the *Named Entity Extraction* procedure receives input document X to extract the named entities. This utilizes an off-the-shelf statistical model (spaCy en_core_web_sm[2]), keeping only those entity chunks corresponding to persons (Line 2), locations (Lines 4 and 3), and nationalities or religious or political groups (Line 5). While looking at the datasets, we realized that it is also relevant to keep the nationality entities, given the frequency they appear surrounding the spans of text that semantically represent political roles (e.g. "*Belgium Prime Minister Alexander De Croo*", "*German Finance Minister Olaf Scholz*", etc). Line 6, calls the function $check\_CAMEO$, querying the CAMEO repositories[3] to map nationalities and more granular locations (e.g., cities, states) into their corresponding locations in countries. If no nationalities or locations are found in CAMEO, then then simply disregard the *loc*, *gpe* or *norp* entity. Algorithm 1 will output the C_X of (person, location) tuples, which serve as inputs for the relation extraction component described in Subsections IV-B, IV-B1 and IV-B2.

B. Relation Extraction Component

The design of our relation extraction component uses transformers-based architectures to leverage transferring learning from existing pre-trained models.

Transfer learning allows the machine to improve the information about a new specific task based on previously learned knowledge from another related task. This implementation

[2]https://spacy.io/models/en

[3]https://github.com/openeventdata/petrarch2/blob/master/petrarch2/data/

Algorithm 1: Named Entity Extraction Procedure

input : News article **X**, NER **model**
output: Collection of (person, location) tuples

```
parsed ← model(X)
person ← parsed.PER
gpe ← parsed.GPE
loc ← parsed.LOC
norp ← parsed.NORP
location ← check_CAMEO(loc + gpe + norp)
C_X ← [Ø]
foreach p in distinct(person) do
    if person p not in CAMEO then
        foreach l in distinct(location) do
            C_X.append((p, l))
return C_X
```

utilizes BERT [38] and RoBERTa [39] with pre-trained models for this specific task/domain. Such technique allows cost savings over human labeling efforts, which is a major constraint.

BERT (Bidirectional Encoder Representations from Transformers) [38] is a language understanding model based the on original transformer structure [36]. The crucial aspect distinguishing BERT from previous models ([49], [50]) is that its multi-layer bidirectional transformer structures enable processing sentences as a whole and learning relationships between words using multi-head attention mechanisms and positional embeddings. Further, the training parallelization property of transformers improves learning efficiency in large datasets, allowing the usage of powerful computational devices to pre-train and fine-tune large-scale language models.

RoBERTa (Robustly Optimized BERT Pretraining Approach) [39] retrains BERT with and improved methodology (optimized hyper-parameters, a modified training task, larger training data, larger mini-batches and learning rates, etc.). Pre-trained models for both architectures are publicly available and can be fine-tuned with a simple additional output layer over the architecture to reach state-of-the-art performance on most of the traditional NLP tasks. Given that they share nearly the same architectures, we assume they are equivalent here.

CoMe-KE implements transformers deep neural networks (BERT and RoBERTa) as part of the RE component to generate contextualized representations of input text. Figure 2a illustrates the usage of DNNs. Let an input document X be composed of a sequence of tokens $[x_0, x_1, \ldots, x_{|X|}]$, where $x_0 = [CLS]$ and $x_{|X|} = [SEP]$ are special tokens. Then feed such tokens into the network, to map the embedding vectors $[E_{[CLS]}, E_1, \ldots, E_{[SEP]}]$ to their corresponding contextualized embedding vectors $T = [T_{[CLS]}, T_1, \ldots, T_{|X|}, T_{[SEP]}]$. These representations $T \in \mathbb{R}^{(|X|+2)\times d}$ of dimension d feed the next layer at the top of the network for a specific task.

Slight modifications of the standard usage of the transformers DNN as contextualized encoders are employed to facilitate their application to RE. Subsections IV-B1 and IV-B2 discuss the approaches utilized to explore the contextualized representations T.

*1) **Distant Supervised (DS) Module**:* Distant supervision [40], [41] is a learning approach widely used in RE, automatically annotating weakly labeled data based on heuristics. The DS module of CoMe-KE learns semantic relations from knowledge available in CAMEO's KB to extract relations between unknown political entities and locations and generate new (actor, role-codes) entries. The DS module implements three steps: (1) constructing **learning data** from the portion of data with known political actors, (2) fine-tuning BERT with learning data, and (3) applying the fine-tuned model to the complementary (**inference dataset**) to discover new entries.

To construct the learning data, simply collect the documents containing at least one known political actor in CAMEO's KB. Formally, for each actor-country pair (p, l) associated through some relation $r \in P$ stored in CAMEO's KB, look for documents containing those entities (both actor and country) in the large unlabeled corpus given as input to form our learning dataset (see, Figure 1). For each input document X, evaluate every actor-country combination (p, l) in X with respect to the possible cases:

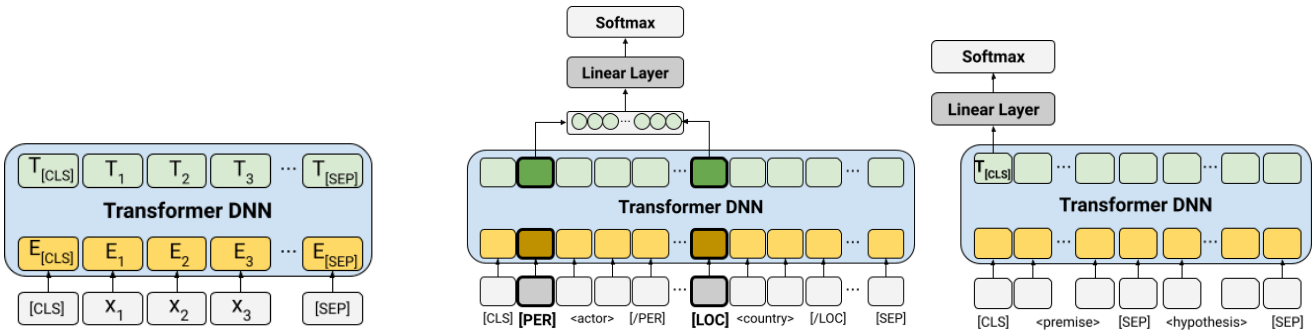
(i) p and l are associated through the semantic relation r and recognized as an entry in CAMEO's KB;
(ii) actor p is found in CAMEO's KB with no semantic relation with the country l;
(iii) actor p is not found in CAMEO's KB.

Next construct the triple (X, p, l) and associate it with the label r for those cases falling in (i) to compose data points in the learning data. For the cases in condition (ii), follow the same procedure for constructing the triple, but we force $r = NOTHING$. Collecting learning data points falling in condition (ii) is essential to learn (during the fine-tuning step) how to identify such cases and avoid generating false-positive entries. Finally, reserve the cases in condition (iii) to be part of the inference dataset, which will be explored to discover new candidate entries and extend CAMEO's KB. The flow toward the bottom in Figure 1 illustrates the procedure.

The constructed learning dataset is the input for fine-tuning the RE model (second step of the DS). As described in Figure 2b, we perform some slight adaptations in the deep transformers contextualized encoder (Figure 2a) for RE purposes. Given data (X, p, l, r) (where r is the step 1 label) in the learning dataset, take $X = [x_0 = [CLS], x_1, \ldots, x_{|X|} = [SEP]]$ as the sequence of input tokens. Let (p_{begin}, p_{end}) and (l_{begin}, l_{end}) be pairs of integers delimiting the index of the tokens for the actor p and country l mentions in X, where $0 < p_{begin} < p_{end} < |X|$ and $0 < l_{begin} < l_{end} < |X|$. Following [42], we adopt **entity marker tokens** $[PER]$, $[/PER]$, $[LOC]$ and $[/LOC]$ as reserved special tokens to mark the begin and end of each entity (PER for actors and LOC for countries). This converts X into $X^{'}$:

$$X^{'} = [[CLS], x_1, \ldots, [PER], x_{pbegin}, \ldots, x_{pend}, [/PER], \ldots, [LOC], x_{lbegin}, \ldots, x_{lend}, [/LOC], \ldots, [SEP]]$$

Following Figure 2b feed $X^{'}$ into a transformer DNN. As output append the contextualized representations for the



(a) Deep Transformer as Contextualized Encoder.

(b) Deep Transformer for Relation Extraction (DS Design).

(c) Deep Transformer for Relation Extraction (ZS).

Fig. 2: Proposed Design for Relation Extraction Component of CoMe-KE.

start states for [PER] (for actors) and [LOC] (for countries) to obtain the **semantic relation representation** $T_{X',p,l} = (T_{[PER]}|T_{[LOC]})$.

Figure 2b shows this semantic relation representation $T_{X',p,l} \in \mathbb{R}^{2d}$ will feed a linear layer represented by the parameter matrix $W \in \mathbb{R}^{|P|\times 2d}$, where P is the set of pre-defined semantic relations (roles). Therefore, for each input (X', p, l), the model will predict a relation $\widetilde{r} \in P$ as follows:

$$\widetilde{r} = argmax\ (W \cdot T_{X',p,l}) \tag{1}$$

where argmax outputs the role in P corresponding to the maximum element's index in softmax's output.

During the fine-tuning step, the RE network learns semantic relation representations such that minimize the cross-entropy loss between the predicted roles $\widetilde{r}$ in Equation 1 and the real label r collected in the construction of learning dataset.

Finally, during inference step, simply output the pair (p=actor, φ=role-code) together with the *confidence score*, computed by the softmax layer with respect to that role $\widetilde{r}$. Note that this step converts both the country l and the predicted role $\widetilde{r}$ associated to the actor p into the role-code $\varphi = alpha\_3(l)|\widetilde{r}$, by appending the ISO alpha-3 code corresponding to country l and role $\widetilde{r}$ (as described in Section III). For example, we would have the outputs in the format: ([Jon Ossoff, USAGOVLEG], score=0.9998.)

*2) **Zero-Shot (ZS) Module**:* This combines *natural language inference* [48] with a *dynamic hypothesis* procedure to devise a ZS-based approach as an alternative for the RE component.

Natural language inference (NLI) models usually take as input two text sequences (a premise and a hypothesis) and returns one of the outcomes: *entailment* (when hypothesis follows from the premise), *contradiction* (when hypothesis contradicts the premise), or *neutral* (when hypothesis neither follows nor contradicts the premise).

The **dynamic hypothesis** procedure proposed here facilitates the usage of NLI models for RE. The procedure receives a set of fixed statements with blanks, which are a dynamic template (designed and tailored by the user), to be filled with the input values — the (p, l) pair.

In this application, for each actor-country pair (p, l), one will have a set of hypotheses of these entities, which will further feed the NLI model built over the transformers network (see, Figure 2c). In contrast to DS, the ZS implementation does not require learning data or fine-tuning. Instead, the ZS implementation simply loads a RoBERTa model pre-trained on the NLI task for direct inference.

Based on existing CAMEO's entries and preliminary analysis run on the dataset utilized on our experiments, the dynamic template and corresponding labels are in Table I.

TABLE I: Template with Dynamic Statements Designed for our Application.

Dynamic Statement	Label (Role)
$\langle p \rangle$ is part of the Ministry of Agriculture of $\langle l \rangle$.	GOVAGR
$\langle p \rangle$ is part of the Ministry of Finance of $\langle l \rangle$.	GOVBUS
$\langle p \rangle$ is part of the Ministry of Development of $\langle l \rangle$.	GOVDEV
$\langle p \rangle$ is part of the Ministry of Education of $\langle l \rangle$.	GOVEDU
$\langle p \rangle$ is a former politician of $\langle l \rangle$.	GOVELI
$\langle p \rangle$ is part of the Ministry of Environment of $\langle l \rangle$.	GOVENV
$\langle p \rangle$ is part of the Ministry of Health of $\langle l \rangle$.	GOVHLH
$\langle p \rangle$ is part of the Ministry of Human Rights of $\langle l \rangle$.	GOVHRI
$\langle p \rangle$ is part of the Ministry of Labor of $\langle l \rangle$.	GOVLAB
$\langle p \rangle$ is Senator of $\langle l \rangle$.	GOVLEG
$\langle p \rangle$ is part of the Ministry of Culture or Communication of $\langle l \rangle$.	GOVMED
$\langle p \rangle$ is part of the Ministry for Refugees of $\langle l \rangle$.	GOVREF
$\langle p \rangle$ is part of the Government of $\langle l \rangle$.	GOV

Thus, the input document X will serve as *premise*, while the $|C_x| \times |P|$ statements provided by our dynamic hypothesis function will serve as the *hypotheses* to input the pre-trained NLI model.

Figure 2c shows how to feed the linear layer $W^{NLI} \in \mathbb{R}^{3\times d}$ with the contextualized representation for the special token [CLS] and apply the softmax function to obtain the dynamic inference (entailment, contradiction or neutral). Thus, for each document-hypothesis pair given as input, obtain the outcome nli using:

$$nli = argmax\ (W^{NLI} \cdot T_{[CLS]}), \tag{2}$$

where argmax gives the NLI outcome for the maximum element's index in the softmax's output.

For each actor p appearing in document X, disregard those statements such that $nli \neq entailment$, keeping only the role associated with the dynamic statement (see Table I) that maximizes the softmax score computed by the NLI model. In other words, assume that the most probable entailment will provide the correct role label.

Finally, as we do in the DS module (Subsection IV-B1), output the pair (p=actor, φ=role-code) with the *confidence score* from the softmax layer for the role hypothesis.

C. End-to-End CoMe-KE

Close this section with CoMe-KE's components altogether in Algorithm 2. As CoMe-KE is designed to work in a pipeline fashion, assume the web-crawling procedure runs in a separate process, obtaining the collection of documents to feed this end-to-end Algorithm 2. Apart from the input corpora, CoMe-KE receives the RE method to be applied (either 'ZS' or 'DS'), the dynamic template to be filled when applying zero-shot method and the threshold that will cut off the low-score entries.

Algorithm 2: CoMe-KE End-to-End Procedure

input : News articles **Docs**, RE **method**, Dynamic Template **template**, Threshold **thresh**
output: Collection of entries p : {role of p, alias of p}

```
1  rank ← {∅}
2  foreach X in Docs do
3      C_X ← Entity_Extraction(X)
4      if method == 'ZS' then
5          hyps ← Dyn_Hypothesis(C_X, template)
6          candidates ← ZS_RE(X, hyps)
7      else
8          candidates ← DS_RE(X, C_X)
9      rank.update(candidates)
10 rank.set_roles()
11 return rank.dedup(thresh)
```

For each document X, first extract the collection of (actors, countries) pairs C_X through the Algorithm 1. If the user opts for the 'ZS' method, then first obtain the hypotheses through the dynamic hypothesis procedure (Subsection IV-B2), which will feed the pretrained NLI model (Figure 2c) together with the document X to obtain the candidate entries. Otherwise, if the user chooses the 'DS', then simply obtain the candidate entries through the distant supervised network.

Note that to apply 'DS' method in Algorithm 2, the user must have fine-tuned the RE model before taking the learning dataset, as discussed in Subsection IV-B1 and illustrated in Figure 1. In contrast, 'ZS' does not require any prior fine-tuning step. As described in Subsections IV-B1 and IV-B2, the RE component outputs a collection of *candidates* in $\langle$(p=actor, φ=role-code), $score\rangle$ format, regardless the RE method.

The variable *rank* in Algorithm 2 is a class implementing a dictionary structure that holds the actor names p as key and a collection of (φ=role-codes, $score$) tuples as values associated to that key. At the end of each loop over the document X, *rank* will be updated with the new *candidates*, as shown in Line 9. After traversing the whole input corpora, *rank* will store all actor names found in all documents as keys. The method $rank.set\_roles()$ called in Line 10 will define the final role-code φ_p for each actor p, by simply choosing that role-code that maximizes the confidence score:

$$\varphi_p = \underset{\varphi_i}{\operatorname{argmax}}\ Pr(\varphi_i|p) \tag{3}$$

where $Pr(\varphi_i|p)$ is given by the $score_{p,i}$ assigned to the role-code $\varphi_{p,i}$ in the list of values [($\varphi_{p,1}$, $score_{p,1}$), ($\varphi_{p,2}$, $score_{p,2}$), ..., ($\varphi_{p,k}$, $score_{p,k}$)] associated to the actor p. As a result, each actor p in *rank* will have only one associated tuple (φ_p, $score_p$).

In Line 11, the *dedup()* method first sorts the *rank*'s entries in ascending order based on the scores attribute and cut off the entries such that score is lower than the threshold *thresh*. Next, for those entries associated to the same role-code, method *dedup()* will deduplicate the actor names based on the resemblance similarity measure [51], keeping together the similar names as aliases. As an example, consider the entries:

```
"Joseph Robinette Biden": "USAGOV"
"Joe Robinette Biden": "USAGOV"
"Joe Biden": "USAGOV"
```

Given that all these entries share the same role-code "USAGOV", the resemblance similarity measure will be able to identify the similarity in the actor names and deduplicate the three entries into the following one, which will be the output format for Algorithm 2:

```
"Joseph Robinette Biden": {"role": "USAGOV",
"aliases": "Joe Biden", "Joe Robinette Biden"}
```

V. Dataset Description

To perform a comprehensive analysis and evaluate CoMe-KE's knowledge extraction efficiency, two datasets based on relevant political and social science topics are employed.

Conflict and Mediation Dataset: We constructed this domain-specific corpus from three types of sources. First, we crawled newswire texts from various world-wide news agencies, and carefully pre-processed and filtered out out-of-domain news based on the metadata information. Second, we collected reports from Amnesty International and Human Rights Watch. These sources contain rich information on political events for conflict research, thus improving the scope of our coverage. Finally, we also considered Wikipedia as an additional source to enrich the diversity of our corpora. Therefore, we queried [52] and curated a collection of relevant articles from an 18 GB Wikipedia dump released on March 20, 2021.[4]

Table II shows the crawled sources grouped by their major region and the overall number of documents extracted. Given the large number of sources and interest in evaluating the performance of the framework on multiple sources, the experiments in Section VI analyze the results by region.

[4]https://dumps.wikimedia.org/

TABLE II: Conflict and Mediation Data (Sources and Sizes).

Region	Sources	# of Docs
Africa (AF)	AllAfrica	85,361
LatinAmerica(LA)	EFE	5,735
Asia (AS)	CNA, IndiaTimes, TheNewsIntl, Xinhua	542,420
Australia (AU)	TheConversation	11,801
Europe (EU)	AFP, BBC, DW, France24, Guardian, Reuters	1,181,313
MiddleEast (ME)	Aljazeera	73,116
US (US)	ABC, CNN, HRW, LATimes, NBC, NPR, NYP, NYT, WaPo, WSJ, USAToday, USNews	1,085,541
Global (GL)	Wikipedia, Amnesty	162,852

For each story in the Table II data, keep the known political actors (from CAMEO's KBs) appearing in this story with their corresponding role-codes as ground-truth for the experiments. Table III shows the total frequency of known political actors appearing in the documents per political role (in the left side of the bar) and the distinct number of known actors (in the right side of the bar). Given the space limitation, the figures show only for three groups (EU, GL and US), but the remaining groups follow the same unbalanced distribution — highly concentrated on GOVELI and GOV roles.

TABLE III: Distribution of Roles for EU, GL and US Corpora (Total Count / Distinct).

Roles	Region		
	EU	GL	US
GOVAGR	1,556 / 210	86 / 68	190 / 77
GOVBUS	5,137 / 326	223 / 120	726 / 104
GOVDEV	692 / 109	15 / 12	40 / 25
GOVEDU	1,014 / 197	124 / 90	183 / 61
GOVELI	258,484 / 1,661	17,508 / 1,612	250,162 / 1,551
GOVENV	1,344 / 140	54 / 42	181 / 50
GOVHLH	3,767 / 257	168 / 105	427 / 110
GOVHRI	383 / 23	14 / 11	58 / 11
GOVLAB	381 / 76	24 / 19	107 / 39
GOVLEG	17,228 / 42	1,863 / 61	52,789 / 71
GOVMED	1,428 / 235	107 / 77	251 / 88
GOVREF	17 / 5	1 / 1	0 / 0
GOV	2,986,570 / 6,950	376,790 / 7,274	306,2291 / 5,829

Organized Crime Dataset: An additional corpus on *organized crime* was collected, another relevant sub-domain in conflict and social sciences but not part of CAMEO. This dataset (Subsection VI-C) demonstrates CoMe-KE's flexibility to explore a schema different than CAMEO's, and potentially extending its repositories to other actors and regions involving political affairs. This real-world dataset consists of news articles reporting organized criminal activity involving gangs, cartels and other non-state armed actors operating in Latin America and the Caribbean. The data were crawled from the *Insight Crime* and contain $13,236$ documents from July 2004 to March 2020.[5]

VI. Experiments

This section describes the computational setup for the experiments and analyze two case studies by applying CoMe-KE for knowledge extraction. Subsection VI-B shows CoMe-KE capabilities for extending CAMEO's repositories following the same (political actor, role-code) schema, traditionally used for event data coding for political and social conflict and mediation. Subsection VI-C, presents a second case study analyzing CoMe-KE's efficiency for potentially extending CAMEO's repositories with actors participating in *organized crime* actions, a category of conflict not currently in CAMEO.

A. *Setup*

We used a computer with NVIDIA RTX 8000 GPU to conduct all the experiments. We utilized off-the-shelf model spaCy for named entity extraction in Algorithm 1. For the DS module, we fine-tuned the BERT large uncased on the constructed learning dataset for 10 epochs with an Adam optimizer, a batch size of 32, a maximum sequence length of 512, and a learning rate of 7e-5.[6] For the ZS module, we used a ready-made RoBERTa large uncased (fine-tuned for NLI task on the MultiNLI dataset [53]).[7] We set *thresh* to 0.3 as cut-off threshold (in Algorithm 2) and utilized exactly the same *dynamic template* depicted in Table I.

To best of our knowledge, the only baseline implemented for extracting political entities in CAMEO-like schema is RePAIR [29]. Therefore, we utilize this as our baseline and keep the same hyper-parameters found by the authors.

B. *Case Study 1: Exploring existing schema in Conflict and Mediation Domain*

The experiments performed in this subsection are based on *Conflict and Mediation Domain* dataset split by regions, (see, Table II). As discussed in Section V, for each document in the corpora, we have as ground truth the known political actors (existing in CAMEO) appearing in the text, with their corresponding role-codes. Since the ground truth labels were assigned based only on CAMEO records, one cannot precisely compute precision. Note that, there will always be two possibilities for a given actor extracted through CoMe-KE, not belong to the ground-truth set: (i) it is not a relevant political actor, and (ii) it is a relevant political actor but is not in the CAMEO repository yet.

Therefore, experiments in this section consider recall-based measures to evaluate the models' performance capturing and recognizing the known existing actors on. We measure recall in the following three different ways:

- **(A&C)-REC** accounts for correctness only if the actor and country if the role-code were correctly retrieved.
- **(A&R)-REC** accounts for correctness only if the actor and political role if the role-code were correctly retrieved.
- **FULL-REC** accounts for correctness only if the actor and the whole role-code were correctly retrieved.

In this experiment, we randomly select a percentage (denoted by **excl.%**) of the distinct political actors appearing in the corpora to represent the **testing set**. The documents containing at least one actor belonging to testing set are input to our models, which should be able to capture such known

[5]https://www.insightcrime.org

[6]https://huggingface.co/bert-large-uncased

[7]https://huggingface.co/roberta-large-mnli

actors and their corresponding roles. We randomly select the testing set for 10 times in each experiment and average the recall measures. Note that actors belonging to the testing set from the CAMEO dictionaries and excluded in the experiments (reason why we call excl.%).

We design the analyses based on *excl.%* and partial testing set because the baseline RePAIR applies a label propagation-based mechanism, depending on known actors with their role-codes to propagate the labels to the potential new actors. Therefore, for the purpose of experimenting RePAIR, we keep the complementary portion of the testing set $(100\% - excl.\%)$ as known actors which will propagate their labels to the potential new actors in testing set.

For experiments with DS, follow the steps in Subsection IV-B1: select all the documents in the corpora containing at least one actor in testing set to obtain the inference set, while the remaining documents will form the learning set. Then fine-tune BERT based model for RE over the learning set and compute recall on the inference set.

Table IV reports the performance of method RePAIR (ReP), CoMe-KE with the distant supervised (DS) and zero-shot (ZS) approaches on retrieving the political actors for the regional corpora. For this experiment, we set excl.% to 20%, meaning that our testing set is composed by 20% of the distinct actors appearing in each regions' corpora.

TABLE IV: Recall on Testing Set per Location (excl%: 20%).

Region	(A&C)-REC			(A&R)-REC			FULL-REC		
	ReP	DS	ZS	ReP	DS	ZS	ReP	DS	ZS
AF	34.0	47.1	**66.8**	26.9	50.4	**60.4**	16.7	34.9	**53.5**
LA	33.6	52.2	**62.8**	25.3	57.6	**60.5**	15.4	39.2	**53.2**
AS	50.6	62.2	**74.8**	34.7	**59.3**	58.7	26.5	44.5	**53.7**
AU	42.8	52.1	**64.6**	30.3	49.4	**63.0**	21.4	37.1	**56.6**
EU	51.7	63.3	**77.6**	35.0	61.6	**63.2**	26.4	45.6	**57.1**
ME	43.2	57.4	**71.1**	29.2	58.4	**66.7**	19.3	40.4	**58.4**
US	38.4	54.7	**72.1**	25.1	50.6	**56.7**	16.1	34.2	**51.0**
GL	39.9	53.8	**54.1**	23.2	**46.1**	44.8	19.2	35.3	**40.4**

Results in **bold** font indicate the best performing model.

Overall, CoMe-KE with zero-shot method (ZS) presents the best results for all regions in both (A&C)-REC and FULL-REC metrics, with large superiority in (A&R)-REC. Both CoMe-KE methods outperform RePAIR in all the regions for all the measures, showing more than the double of RePAIR's FULL-REC for most of the regions. CoMe-KE's performance improvements over RePAIR are statistically significant at the 0.001 confidence level (based on t-test).

Note that Table IV analyzes only the recall-based measures in micro-averaged fashion. Table V, reports the models' performance by political roles on the whole dataset (all regions together). This analysis shows the efficiency of both CoMe-KE variants retrieving political actors for all the roles. Apart from CoMe-KE's superiority in overall macro FULL-REC, note that DS approach obtained reasonable performance for all roles, except by GOVHRI and GOVREF. We conjecture that the low performance is explained by the fact that the extremely small number of actors on those roles (see Table III) is not enough for the network to correctly learn the concepts during the fine-tuning process. On the other hand, ZS obtained lower performance on those roles whose domains tend to be more heterogeneous. For example, GOVBUS can be either a ministry of finance or a national central bank president while a GOVMED can be either a ministry of culture, minister of communication or government press officer. We suspect that these roles require a more precise design for dynamic statements than the ones implemented in Table I.

TABLE V: Recall on Testing Set by Role (excl%: 20%).

Role	FULL-REC		
	ReP	DS	ZS
GOVAGR	2.4	41.0	**49.2**
GOVBUS	4.9	**32.7**	13.2
GOVDEV	1.4	22.1	**33.1**
GOVEDU	2.7	**47.6**	41.3
GOVELI	14.6	29.1	**58.4**
GOVENV	2.3	40.8	**44.2**
GOVHLH	2.7	**55.0**	51.9
GOVHRI	0.0	16.2	**33.1**
GOVLAB	1.6	**37.9**	34.4
GOVLEG	8.3	**34.0**	11.6
GOVMED	1.6	**36.7**	10.8
GOVREF	0.0	8.3	**29.4**
GOV	30.9	43.7	**57.5**
MACRO	5.6	34.2	**36.0**

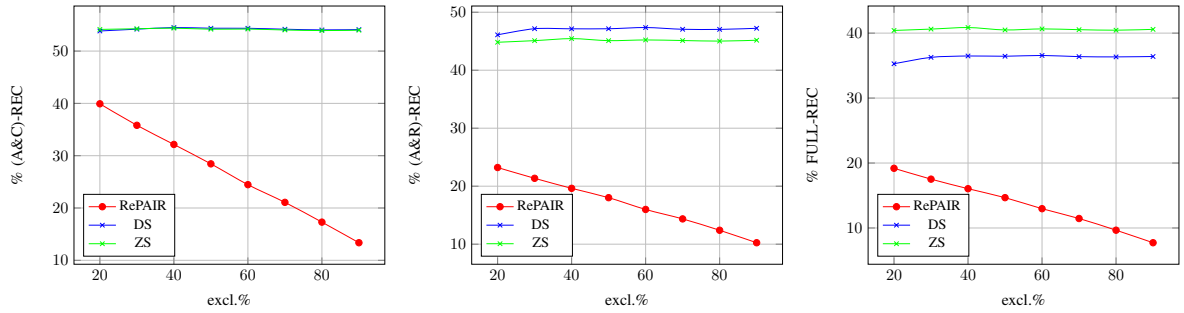
Results in **bold** font indicate the best performing model.

Table V shows an additional advantage: CoMe-KE is not sensitive to the existing known actors. While RePAIR depends on the existing known entries to propagate and assign the correct role-codes, CoMe-KE does not. It is notable in Table V that RePAIR presented reasonable recall only for those dominant roles containing the largest number of actors appearing in the corpora (GOV, GOVELI and GOVLEG - see Table III), while both CoMe-KE methods show good performance for most of the political roles. To further test the findings in this direction, the experiment in Figure 3 over Global (GL) data, presents the recall measures while decreasing the proportion of known entries. Increasing the excl% (increase the testing data and decrease the proportion of known entries), drastically decreases the recall for RePAIR, while it stays stable for CoMe-KE.

Finally, we estimate the precision of our models on retrieving political actors and their role-codes by manually evaluating the output records. For this exercise, we simulate a real application of our models, by running RePAIR and CoMe-KE (Algorithm 2) over Global (GL) corpora to retrieve new actors not belonging to CAMEO.

Given that CoMe-KE and RePAIR have their own methods for sorting the output recommendations, we evaluate their ranking quality and the precision on the top ranked actor recommendations. For that purpose, we submitted the top 500 (actor, role-code) pairs output from each model for manual validation by political science experts. We compute precision in the same fashion as we previously did for recall, and also report the distinct number of role-codes among the top 500 retrieved actors, in Table VI.

Although RePAIR presents the best numbers for (A&R)-PREC, all of its top 500 recommended actors were from USA with role-codes GOV or GOVELI. On the other hand, both



(a) (A&C)-REC (b) (A&R)-REC (c) FULL-REC

Fig. 3: Recall on Global (GL) dataset by excl%

TABLE VI: Precision and Distinct Role-codes on Top 500 Recommended Actors.

Model	(A&C)-PREC	(A&R)-PREC	FULL-PREC	Distinct #
ReP	55.0	**91.4**	53.0	2
DS	76.7	90.2	82.9	**130**
ZS	**91.6**	88.4	**87.2**	108

Results in **bold** font indicate the best performing model.

CoME-KE configurations covered a larger range of political actors. Lastly, both CoMe-KE configurations significantly outperform RePAIR in (A&R)-PREC and FULL-PREC.

C. *Case Study 2: Exploring new schema on Organized Crime Domain*

A second study case explores an *organized crime* corpus to extend CAMEO's KBs with criminal actors, which is a category of conflict currently not covered by CAMEO. Instead of following the traditional CAMEO schema based on role-codes, our goal is to extract criminal entities, the country where they are rooted, and most frequent types of crimes they commit, following simple ontology designed in Figure 4.

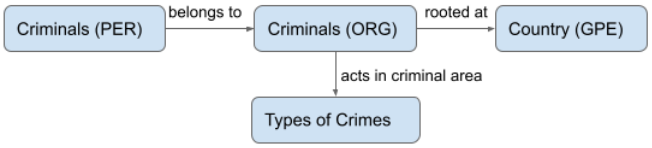


Fig. 4: Design of the Schema Explored for Extracting Knowledge on Organized Crime Corpus.

To extract the knowledge in such s format, we perform some slight modifications in Algorithm 1 to also extract ORG named entities (for criminal organizations) and output pairs in the format (person, organization) and (organization, country). We also tailor the dynamic templates to capture the semantic relations between criminals (PER) and gangs (ORG), and gangs (ORG) and countries (GPE, LOC). Note that for exploring this new domain one can neither RePAIR nor the DS method, since there is no background KB with known criminals and gangs from which to learn it. We submitted the top 200 results of each relation to our political science experts, obtaining interesting results. Some of the knowledge extracted from these top recommendations are reported in Table VII, which shows the criminal organizations with some of their members, country where they are rooted and top 3 most frequent types of crime.

Based on the feedback provided by our experts, there is 51% of precision for the semantic relation *belongs_to (criminal, organization)*, and 65% for *rooted_at (organization, country)*. To retrieve types of crimes, the dynamic template covers the 18 most common types of crimes reported in the source corpus. Given the complexity on evaluating common crime categories committed per organization, we omit the performance for this relation.

The low precision obtained for this exercise was certainly impacted by the noise introduced by Spanish language sources. Since the corpus extracted covers the Latin America region, most of the names of the gangs, criminal organizations and general citations about criminal aliases and places are mentioned in Spanish. Another interesting detail discovered while exploring this domain is the fact that most of criminals are usually known by aliases, which commonly appear more often in the news than their own real names. Therefore, exploring multilingual pre-trained models as well as eventually incorporating a co-reference resolution in CoMe-KE (to associate aliases and criminal names) are part of our next steps.

VII. Conclusions and Future Work

This paper introduced the CoMe-KE framework to extend automatically knowledge bases in the Conflict and Mediation domain. It implements an off-the-shelf model to extract pairs of entities, and two modules for semantic relation extraction. While a distant supervised module learns from the existing knowledge available in the KB, the zero-shot module explores and innovative *dynamic hypothesis* approach to extract knowledge based on a template tailored by the end-user. Both CoMe-KE approaches require no labeled data to obtain excellent results.

The experiments show that CoMe-KE significantly outperforms the existing baseline on capturing new political entities. Furthermore, we demonstrate the flexibility of CoMe-KE through ZS module on exploring other social science's sub-domains, such as organized crime.

TABLE VII: Extracted Knowledge Sample from Organized Crime Corpora

Organizations	Members	Country	Types Of Crimes
Red Command	Marcelo Pinheiro Veiga, Márcio dos Santos Nepomuceno, Elias Pereira da Silva, Alexander Mendes da Silva,Luiz Fernando da Costa, Wagner Santulho	BRA	contraband, eco-trafficking, microtrafficking
Sinaloa Cartel	Jesus Martinez Espinosa, Joaquin El Chapo Guzman, Jorge Milton Cifuentes Villa, Mario Segovia, Jimmy Waine Galliel, Ismael Zambada Garcia, Mario Nunez Meza	MEX	eco-trafficking, contraband, arms-trafficking
Los Zetas	Rogelio Gonzalez Pizana, Alvaro Gomez Sanchez, Daniel Perez Rojas, Jose Guadalupe Reyes Rivera, Arturo Guzmán Decena, Sergio Ricardo Basurto Peña	MEX	eco-trafficking, arms-trafficking, microtrafficking
ELN	Odin Sanchez Montes de Oca, Fredy Moreno Mahecha,Nicolas Rodriguez Bautista, Carlos Germán Velasco Villamizar, Ramiro Vargas, Carlos Arturo Velandia	COL	contraband, eco-trafficking, arms-trafficking
Texis Cartel	José Adán Salazar Umana, Jose Misael Cisneros Rodriguez, Moris Alexander Bercian Machon, Julio Cesar Bonilla Cabrera,Robert Antonio Herrera Hernandez	SLV	contraband, eco-trafficking, corruption
Beltran Leyva	Edgar Valdez Villarreal, Sergio Villarreal Barragan, Hector Beltran Leyva	MEX	contraband, eco-trafficking, arms-trafficking
FARC	Guillermo Leon Saenz, Carlos Ivan Mendes Mesquita, Jefferson Chávez Toro, José Benito Cabrera Cuevas, Anayibe Rojas Valderrama, Arturo Ruiz, Walter Arizala	COL	contraband, eco-trafficking, arms-trafficking
MS-13	Moris Alexander Bercian Manchon, Richard Castillo Salazar, Nelson Alexander Flores, Marco Antonio Sian Chavez, Eduardo Erazo Nolasco, Jose Luis Mendoza Figueroa	SLV	arms-trafficking, contraband, homicides

Future work will focus on improving the entity recognition component by eventually replacing the off-the-shelf model by a BERT-based sequence labeling model and experiment with other PLMs as basis of the network.

ACKNOWLEDGEMENTS

The research was supported in part by NSF awards OAC-1931541, OAC-1828467, DMS-1737978, DGE-2039542, and DGE-1906630, ONR awards N00014-17-1-2995 and N00014-20-1-2738, Army Research Office Contract No. W911NF2110032 and IBM faculty award (Research). The authors are solely responsible for the contents.

REFERENCES

[1] J. Bercovitch and S. S. Gartner, "International conflict mediation," *Abingdon: Routledge*, 2009.

[2] E. E. Azar, "The conflict and peace data bank (COPDAB) project," *Journal of Conflict Resolution*, vol. 24, no. 1, pp. 143–152, 1980.

[3] B. O'Connor, B. M. Stewart, and N. A. Smith, "Learning to extract international relations from political context," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1094–1104, 2013.

[4] A. Hanna, "Mpeds: Automating the generation of protest event data," Available at https://osf.io/preprints/socarxiv/xuqmv (2021/08/07), 2017, unpublished Manuscript.

[5] J. Osorio and A. Reyes, "Supervised event coding from text written in spanish: Introducing eventus id," *Social Science Computer Review*, vol. 35, no. 3, pp. 406–416, 2017.

[6] J. Osorio, A. Reyes, A. Beltrán, and A. Ahmadzai, "Supervised event coding from text written in Arabic: Introducing hadath," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 49–56.

[7] G. Glavaš, F. Nanni, and S. P. Ponzetto, "Cross-lingual classification of topics in political texts," in *Proceedings of the Second Workshop on NLP and Computational Social Science*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 42–46.

[8] J. Beieler, "Generating politically-relevant event data," *In Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 37–42, 2016.

[9] B. Radford, "Multitask models for supervised protest detection in texts," Available at https://arxiv.org/abs/2005.02954 (2021/08/07), 2019, unpublished Manuscript.

[10] E. S. Parolin, S. Salam, L. Khan, P. Brandt, and J. Holmes, "Automated verbal-pattern extraction from political news articles using CAMEO event coding ontology," *International Conference on Intelligent Data and Security*, pp. 258–266, 2019.

[11] E. S. Parolin, L. Khan, J. Osorio, V. D'Orazio, P. Brandt, and J. Holmes, "HANKE: Hierarchical Attention Networks for Knowledge Extraction in political science domain," *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2020.

[12] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing ELMo and DistilBERT on socio-political news classification," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 9–18.

[13] F. Olsson, M. Sahlgren, F. ben Abdesslem, A. Ekgren, and K. Eck, "Text categorization for conflict event annotation," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 19–25.

[14] F. K. Örs, S. Yeniterzi, and R. Yeniterzi, "Event clustering within news articles," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 63–68.

[15] B. Radford, "Seeing the forest and the trees: Detection and cross-document coreference resolution of militarized interstate disputes," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 35–41.

[16] E. S. Parolin, L. Khan, J. Osorio, P. Brandt, V. D'Orazio, and J. Holmes, "3M-transformers for event coding on organized crime domain," *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2021.

[17] C. Norris, P. Schrodt, and J. Beieler, "PETRARCH2: Another event coding program," *Journal of Open Source Software*, vol. 2, no. 9, p. 133, 2017. [Online]. Available: https://doi.org/10.21105/joss.00133

[18] S. Salam, P. Brandty, J. Holmesy, and L. Khan, "Distributed framework for political event coding in real-time," in *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*. IEEE, 2018, pp. 266–273.

[19] S. Salam, L. Khan, A. El-Ghamry, P. Brandt, J. Holmes, V. D'Orazio, and J. Osorio, "Automatic event coding framework for spanish political news articles," in *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2020, pp. 246–253.

[20] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 697–706.

[21] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia," *Artificial Intelligence*, vol. 194, pp. 28–61, 2013.

[22] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*. Springer, 2007, pp. 722–735.

[23] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 2008, pp. 1247–1250.

[24] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 601–610.

[25] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[26] N. Nakashole, M. Theobald, and G. Weikum, "Scalable knowledge harvesting with high precision and high recall," in *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, 2011, pp. 227–236.

[27] N. Nakashole, G. Weikum, and F. Suchanek, "Patty: A taxonomy of relational patterns with semantic types," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1135–1145.

[28] F. Niu, "Web-scale knowledge-base construction via statistical inference and learning," Ph.D. dissertation, The University of Wisconsin-Madison, 2012.

[29] M. Solaimani, S. Salam, L. Khan, P. T. Brandt, and V. D'Orazio, "RePAIR: Recommend political actors in real-time from news websites," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1333–1340.

[30] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," *arXiv preprint arXiv:1906.05317*, 2019.

[31] O. Etzioni, A. Fader, J. Christensen, S. Soderland *et al.*, "Open information extraction: The second generation," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[32] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1535–1545.

[33] M. Schmitz, S. Soderland, R. Bart, O. Etzioni *et al.*, "Open language learning for information extraction," in *Proceedings of the 2012 joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 523–534.

[34] J. Fan, D. Ferrucci, D. Gondek, and A. Kalyanpur, "Prismatic: Inducing knowledge from a large scale lexicalized relation resource," in *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, 2010, pp. 122–127.

[35] L. Cui, F. Wei, and M. Zhou, "Neural open information extraction," *arXiv preprint arXiv:1805.04270*, 2018.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, unpublished Manuscript.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019, unpublished Manuscript.

[40] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, 2009.

[41] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 541–550, 2011.

[42] L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," *arXiv preprint arXiv:1906.03158*, 2019.

[43] P. K. Pushp and M. M. Srivastava, "Train once, test anywhere: Zero-shot learning for text classification," *arXiv preprint arXiv:1712.05972*, 2017.

[44] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, "Zero-shot relation extraction via reading comprehension," *ArXiv*, vol. abs/1706.04115, 2017.

[45] A. Obamuyide and A. Vlachos, "Zero-shot relation classification as textual entailment," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 72–78. [Online]. Available: https://aclanthology.org/W18-5511

[46] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," *arXiv preprint arXiv:1909.00161*, 2019.

[47] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[48] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[49] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *NAACL*, 2018.

[50] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

[51] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997, pp. 21–29.

[52] M. Manske, "Petscan," https://petscan.wmflabs.org/, 2019, accessed: 2021-03-20.

[53] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101