Systematizing News Reports on Conflict Incidences

Team

- Cristina Mac Gregor
- Gonzalo Pons
- Darshan Sumant

Data

ACLED (Armed Conflict Location & Event Data), available at https://www.acleddata.com/ is a comprehensive dataset that has as units of observation confrontations that can be catalogued as a conflict incidence under four main categories: battles, violence against civilians, riots/protests, or remote violence. It includes information from countries in Africa (since 1997), countries in Asia (depending on the country, some since 2010, and some since 2016), countries in the Middle East (reports since 2016) and countries in Europe (since 2018). The most updated version of the data has information covering until April 6, 2019.

Some relevant information provided for each event includes the following (see appendix for table with complete list of variables):

1. Country and locality where the event took place.
2. Actors involved (Actor 1 and Actor 2)
3. Classification of the actors into one of 7 categories (e.g. State forces, Rebel groups, Rioters, Civilians)
4. Number of fatalities.
5. A category of the interaction between both actors. This is computed by concatenating the classification of actors (e.g. SOLE MILITARY ACTION, MILITARY VERSUS REBELS, POLITICAL MILITIA VERSUS CIVILIANS).
6. Location of the event, pinned down to latitude-longitude coordinates.
7. Text snippet with more information from the source.

Some examples of how the text looks like are the following:

[177] *'Troops have shot dead a commander of Algerias most radical guerrilla faction who had been sought for scores of killings, a pro-government newspaper said. LAuthentique said Hamou Eulmi, known by his nom de guerre Zinedine, was killed Tuesday near a mos'*

[192] *'25 March 1999 The Globe and Mail Two girls aged 2 and 3 and a woman were among nine victims who had their throats cut overnight by suspected Islamic extremists at an isolated farm south of Algiers, security services said yesterday.'*

[1123] *'One militant was also killed Thursday in a clash with security forces in Beni Beshir, close to Skikda.'*

[19293] *'Violent fight between Seleka rebels and rioters from the Gobongo in Bangui'*
[112212] *'Al Shabaab attack a military base. No reported casualties.'*

To create and update this dataset, analysts first collect information from secondary sources, such as news articles or reports. This information is then coded by a first reviewer. Two more reviewers check this information to ensure quality. ACLED provides detailed guidelines for each of these steps.

Objectives

1. **Goal 1: Identify actors and whether there was a fatality:** This section of our project is partly a *name entity recognition problem.* We aim at identifying which are the contending parts in a conflict given the snippet of text we are given. An advantage of the dataset that we are working with is that we have labelled data which allows us to train a model through supervised methods. Moreover, we also aim at inferring from the text if there were 0 or a positive number of killings in the event.

2. **Goal 2: Perform topic analysis to explore additional patterns in the reported events:** Through *unsupervised methods* we aim at identifying underlying patterns from the text that could help categorize the data in other ways than the existing ones. In particular we believe we could approach this section through *topic modelling.*

3. **Goal 3: Identify patterns in the type of violence perpetrated, and the direction in which the attack was inflicted.** We aim at using *relation extraction* methods for categorizing the different ways in which Actor1 and Actor2 interact, as well as the direction of the interaction. The first step is identifying the verb of the action that defines the relationship. Due to the nature of the data we're working with we expect them to be along the lines of "killed", "bombed", "occupied" and other violent actions. A second step we want to take is to understand the directionality of this relation. Was it A that attacked B or B that attacked A?

Related work

We found a series of articles that worked with *name entity recognition problem, topic modelling* and *relation extraction*. Even though we didn't find papers that solve the exact same problems, we focus here in three papers that solve a similar problem in comparable contexts.

1. ***Source 1: Identifying civilians killed by police with distantly supervised entity-event extraction. Katherine A. Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. Proceedings of EMNLP 2017.***

Abstract: The authors aim to extract names of persons who have been killed by the police from news text, using EM-based distant supervision with logistic regression and convolutional neural network classifiers. They describe this task as combining *event extraction* and *relation extraction,* by describing events, their arguments, and the semantic relation between entities.

Data: The news documents were collected from continually querying Google News throughout 2016 with using keywords related to police and fatality. This searches were restricted mostly to the US.

Pre-processing: They use spaCy NLP package to segment sentences and extract entity mentions. This included splitting the data in to sentences, removing duplicates of sentences, removing sentences with less than 5 tokens or more than 200, among other data cleansing tasks. Regarding entity mentions, they extract mentions identified as "persons" by spaCy's named entity recognizer and that have a (first name, last name) pair as analyzed by the HAPNIS rule-based name parser. To prepare sentences for modeling, candidate mention spans are collapsed to a special "Target" symbol, while other names are mapped to a different "Person" symbol. Additionally, to improve precision, the data was filtered to include only sentences with both a police and fatality keyword.

Learning models: both feature-based logistic regression and convolutional neural networks combined with both "hard" distant label training and "soft" Expected Maximization (EM) joint training. For the CNN approach a stochastic gradient descent for the negative expected log-likelihood is performed using pre-trained word embeddings for initialization.

Metrics: Area under precision recall curve (AUPRC) and F1 scores.

Evaluation of results: The models predict entity-level labels and the authors compare these with a dataset of Fatal Encounters which has more than 18,000 entries of victims' name, age, gender, race, in addition to location, cause and date of death. This comparison is done with two *distant supervision training* paradigmas ("hard" and "soft").

2. ***Source 2: Learning to Extract International Relations from Political Context. Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. Proceedings of ACL 2013.***

Abstract: The authors present an unsupervised approach to event extraction, through a probabilistic model that infers latent frames and a representation of the relationship between political actor pairs. It aims to produce tuples with the form (*source (actor)*, *receiver (actor)*, *timestamp*, *predicate path (action)*)

Data: 6.5 million newswire articles from the English Gigaword 4th edition (1994–2008, Parker et al., 2009), and a sample from the New York Times Annotated Corpus.

Pre-processing: syntactic pre-processing. CoreNLP is used for POS-tag and parsing the articles. Name Entity Recognition (NER) is done deterministically by finding instances of country names from the CountryInfo.txt dictionary from TABARI. To identify the verbs and their direction, they use the "CCprocessed" version of the Stanford Dependencies (de Marneffe and Manning, 2008). Verb paths are identified by looking at the shortest dependency path between two mentions in a sentence.

Learning models: Logistic normal topic model, including latent temporal smoothing on the political context prior.

Metrics & Evaluation of results: The authors compare the automatically learned verb classes to pre-existing verb patterns from TABARI (open-source rule-based event extraction system), and demonstrate correlation to the Militarized Interstate Dispute (MID) dataset. The metric they use is the Area Under the Curve metric.

3. ***Source 3: Content-driven, unsupervised clustering of news articles through multiscale graph partitioning. M. Tarik Altuncu, Sophia N. Yaliraki, Mauricio and Barahona. Data Science, Journalism & Media workshop 2018.***

Abstract: The authors showcase an unsupervised approach to content-clustering, at different levels of resolution, by using a recent deep neural network text analysis methodology (Doc2vec) that represents text in a vector form and then applies a multi-scale community detection method (Markov Stability) to partition a similarity graph of document vectors.

Data: The authors use a corpus of 9,021 news articles published by Vox Media during the whole of 2016. The articles correspond to a wide range of topics from politics, to sport, to human interest, with a clear focus on the US. News about the 2016 US Presidential election, Brexit, and the Rio Olympics feature heavily in the corpus.

Separately, the base Doc2vec model is trained using a corpus of 5.4 million Wikipedia articles. This model trained on the Wikipedia corpus is then applied to the Vox Media corpus.

Pre-processing: Tokenization, with removal of stop words, and normalization using NLTK module stemming methods. Meaningless wrapper sentences such as document header, footer, signatures, annotations, etc. were removed with a threshold of 2+ frequency on the sentence tokens.

Learning models: The base Doc2vec model is trained using a corpus of 5.4 million articles from Wikipedia. This is done to ensure a good representation of the general text encountered in news articles. The model thus trained, is then applied to the corpus of 9021 Vox Media articles from 2016 (after the pre-processing to tokenize, clean wrappers, etc.) to create a 300-dimensional paragraph vector for each of the 9021 articles. This paragraph vector encapsulates a variety of semantic & syntactic characteristics.

Pairwise cosine similarity between all possible pairs ($^{9021}C_2$) of articles is obtained, and for the entire dataset, a min spanning tree (MST) is generated through the MST-kNN method with k = 13 (nodes are articles, and the weighted edges represent similarities between the nodes) which is called the similarity graph $\boldsymbol{G_S}$. a Markov Stability (MS) is applied to $\boldsymbol{G_S}$ in order to extract the multi-scale community structure intrinsic to the graph. MS defines hard clusters based on partition extracted from the similarity graph.

Metrics & Evaluation of results: The quality of clusters generated is evaluated using two alternative approaches - (1) Intrinsic Topic Coherence is measured using the pointwise mutual information (PMI) based on the co-occurrence of the 15 most common words. (2) To compare similarity between partitions of the similarity graph, the authors use normalized mutual information (NMI) based on mutual information between, and the individual entropies of the two cluster assignments.

Other references we have found:

4. *Source 4: Analyzing Entities and Topics in News Articles Using Statistical Topic Models*
5. *Source 5: Collective Cross-Document Relation Extraction Without Labelled Data*

Plan of action

We have numbered our goals for the project in order of feasibility. We prioritize having a strong, well-modeled and highly predictive model for goal 1 before moving forward and working on the subsequent goals. Our plan of action follows the same order: once we have finished working on our first goal we will address goals 2 and 3.

Pre-processing the data: The dataset is overall very clean and structured, and it does not need a lot of work on our behalf. We will drop observations with no input for the text column, or with no labels for the ACTOR1 and ACTOR2 columns.

Supervised learning (for goal #1 and #3)

- Tokenize and vectorize the text input.

- Using existing software (either GloVe or word2vec) we initialize word embeddings from the text.
- **Name entity recognition (Goal #1)**
    - Use SpaCy to perform regular Name Entity Recognition. We want to keep those identified as a GPE (geopolitical entity)
    - Attempt to perform also a bidirectional LSTM (neural network approach) through Tensorflow and Keras (reference)
- **Relation extraction (Goal #3)**
    - We follow the work done by O'connor et al and Keith et al. to identify the relationship between actors.
    - After Implementing goal #1 and getting the outcome for the Actors prediction we will apply the *Stanford Dependencies* software to try to detect the relationship between actors and its direction.
    - Moreover, following Keith, et al:
        - Perform the neural network analysis, with the initialized word embeddings. While the authors use Convolutional Neural Networks, we might instead try Recurrent Neural Networks.
        - Construct hand-crafted features from text. Authors include two sets of features:
            - Syntactic dependencies: 1) Length 3 dependency paths that include TARGET: word, POS (*part of speech*), dependent label; 2) Length 3 dependency paths that include TARGET: word and dependent label; 3) Length 3 dependency paths that include TARGET: word and POS; 4) All length 2 dependency paths with word, POS (*part of speech*), dep. Labels. The tools for building these features would be Stanford's dependencies software.
            - N-gram features: 1) n-grams length 1, 2, 3; 2) n-grams length 1, 2, 3 plus POS (*part of speech*) tags; 3) n-grams length 1, 2, 3 plus directionality and position from TARGET (and in our case TARGET will be ACTOR1 and ACTOR2); 4) Concatenated POS (*part of speech*) tags of 5-word window centered on TARGET; 5) Word and POS (*part of speech*) tags for 5-word window centered on TARGET
        - Perform traditional supervised machine learning methods (Logistic Regression, as the authors do, and Naive Bayes) using SciKit Learn to learn the matrix W and perform predictions.
- **For each problem (name entity recognition and relation extraction)**
    - We will randomly separate the data into train/validation and test sets using a split of 80-20 percent. Choose best model according to implementation on the train/validation set through "hold-out" (randomly splitting the train/validation data into specific train and validation sets) or k-fold cross-validation (using SciKit Learn as well).
    - Evaluate best model on test set.

Unsupervised learning section (Goal #2).

- Tokenize
- Remove stopwords
- Create a term-frequency vectorized representation of the text
- Implement Latent Dirichlet Allocation (LDA)
- Map the newfound topics to events
- Plot geographically and other exploratory analysis. How are these topics related to the rest of the information we have about the conflict, such as country of origin and actors involved?
- Alternatively, use the Doc2vec clustering model with Markov Stability.

Resources:

- Parsers and part of speech tagging:
    - SpaCy and Standord CoreNLP
- Neural network approaches:
    - PyTorch; Keras; TensorFlow
- Machine learning implementation:
    - SciKitLearn
- Dependency analysis:
    - Stanford Dependencies:
    - https://nlp.stanford.edu/software/stanford-dependencies.htm
    - As stated on the website, “Stanford dependencies provides a representation of grammatical relations between words in a sentence. They have been designed to be easily understood and effectively used by people who want to extract textual relations. Stanford dependencies (SD) are triplets: name of the relation, governor and dependent.“

Mid-quarter expectation: For the mid-quarter presentation we aim at having completed a first and complete pipeline that processes the data for our first goal. This would include building all features and a rough pipeline both for the supervised machine learning section of the first goal and the neural networks approach. While we don’t know how predictive our model will be at this stage, we will have information to decide the best way to proceed with the project: If it’s best to continue fine-tuning our model for goal #1 or if we can move on and address goals #2 and #3.

Final presentation expectation: We will have completed goal #1 and goal #2. Time permitted, we hope to have addressed goal #3 as well.

Evaluation of work

For goal 1: There are two different sub-classification problems in our first goal. We want to:

1. Identify both actors of the conflict: Our y_hat is a tuple including two actors, and our prediction will be correct if those two actors are the same as the labels we have from ACTOR1 and ACTOR2 variables, regardless of the order in which they appear. If at least one is different, our prediction will be incorrect. We will generate a binary outcome variable through which we will be able to calculate precision, recall, and maximize Area Under Curve (AUC) measure.
2. Identify if a conflict resulted in fatalities: Similarly than the previous sub-classification problem, we will generate a binary outcome variable, taking the value of 1 if our prediction is correct and 0 if it is not, and we will maximize AUC.

As mentioned, given that we have labelled data for these problems, we will randomly split the data into train/validation and test, performing either "hold-out" or k-fold cross-validation in the train/validation set, and then evaluate the performance of the best model in the test set.

For goal 2: We will choose the number of topics with indicators such as the Coherence score, as well as how neatly interpretable these are (which will have to be determined manually). Likewise, we will explore these topics in relation to other variables in the dataset, such as the type of actors or locations (reference).

For goal 3: We don't have labelled data on the directionality of the attacks. The main references we found use a method called *distant supervision* in which the relations found are evaluated against an external *gold standard data* that includes most information about the relationships analyzed. We have been looking for an equivalent reference, but have had no luck finding something at the micro level at which we are working. The alternative we have thought about is using another variable that we have in the dataset called *Interaction* which includes a numeric code corresponding to the relationship between types of ACTOR1 and ACTOR2. This variable, however, is not explicit about the directionality of the attacks. We would have to restrict our data to those cases from which we could infer in broad terms the directionality; for example, cases in which the population attacked are civilians. We have identified some definitions of the interactions that allow for this inference. For example:

[37] - POLITICAL MILITIA VERSUS CIVILIANS (e.g. outsourced state repression carried out by pro-government militias; civilian targeting by political militias or unidentified armed groups)

[40] SOLE COMMUNAL MILITIA ACTION (e.g. destruction of property by a communal militia; establishment of a local security militia)

We can restrict our data to those in which we can infer the relationship for modelling goal 3.
Moreover, we can draw a sample from the rest of the data, manually create labels for the relations and its directionality, and follow a bootstrapping approach to train, validate, and scale.
We will further explore ways to evaluate the performance of goal #3.

Caveats

1. There are cases in which the snippets presented in the data don't have the information about the actors even though the Actor sections of the data are labelled (presumingly through inference from unreported sections of the text or broader information about the conflict). This is a caveat that we can't address at the moment, and we will work with the fact that our data is not complete.
2. The text snippets that we have is not necessarily representative of what we would see in a news report. We only have the snippet of the report with information about the main event and the actors involved. Our analysis would only be replicable for text in similar format, such as tweets containing brief descriptions of events and certain keywords.
3. Our analysis does not intend to infer which news are relevant to identify a conflict incidence. The scope of the analysis is limited in the sense that any input given to our model has already been classified as related to a conflict incidence (i.e. our data doesn't have text snippet of events that are not conflict-related).

Team-work division

Instead of dividing different steps of the process, our team is planning on holding bi-weekly (twice a week) work sessions to work together on the different steps of the process. In general terms, we will divide up the functions that we will build but also collaborate on them to ensure clean, parameterized, and replicable code. As mentioned, we will try to work in one goal at a time, building up to the most complex task.

Appendix

Appendix 1: Table of columns in dataset (taken from ACLED resources)

Column Name	Content
ISO	A numeric code for each individual country
EVENT_ID_CNTY	An individual identifier by number and country acronym (updated annually)
EVENT_ID_NO_CNTY	An individual numeric identifier (updated annually)
EVENT_DATE	The day, month and year on which an event took place
YEAR	The year in which an event took place

TIME_PRECISION	A numeric code indicating the level of certainty of the date coded for the event
EVENT_TYPE	The type of event
SUB_EVENT_TYPE	The type of sub-event
ACTOR1	The named actor involved in the event
ASSOC_ACTOR_1	The named actor associated with or identifying ACTOR1
INTER1	A numeric code indicating the type of ACTOR1
ACTOR2	The named actor involved in the event
ASSOC_ACTOR_2	The named actor associated with or identifying ACTOR2
INTER2	A numeric code indicating the type of ACTOR2
INTERACTION	A numeric code indicating the interaction between types of ACTOR1 and ACTOR2
REGION	The region of the world where the event took place
COUNTRY	The country in which the event took place
ADMIN1	The largest sub-national administrative region in which the event took place
ADMIN2	The second largest sub-national administrative region in which the event took place
ADMIN3	The third largest sub-national administrative region in which the event took place
LOCATION	The location in which the event took place

LATITUDE	The latitude of the location
LONGITUDE	The longitude of the location
GEO_PRECISION	A numeric code indicating the level of certainty of the location coded for the event
SOURCE	The source of the event report
SOURCE_SCALE	The scale (local, regional, national, international) of the source
NOTES	A short description of the event
FATALITIES	The number of reported fatalities which occurred during the event
TIMESTAMP	Time stamp
CONTINENT	Continent of the event