Prediction Of Major Events

Project in Artificial Intelligence – 236502

Ibrahiem Assad

204524847

simplex3@campus.technion.ac.il

Bar Albo

315668590

albobar@campus.technion.ac.il

The Problem's Definition:

We divide the project into two parts:

The First Part: **Prediction Of Crises**

In the first part, we will focus on a very small subset of major events - namely crises and predict if such will occur in a particular country at a given time.

We will start with one particular type of crisis: DPC (Domestic Political Crisis)

Given a country C and time t (a specific month and a year); predict the probability of a DPC at time t:

$$DPC_C(t) = ? \qquad OR$$

$$P\left(ev_{DPC_C}(t + \Delta) \,\middle|\, ev_c(t), \dots, ev_c(t-k)\right) = ?$$

This may be expanded to several types of crises which are labeled in the data that we can predict, expending the original problem to prediction of any kind of crisis (Political or International) that can occur in a given country at time t.
Therefore, we will continue with subset of all the kinds of crises which are defined in the dataset:

Given a country C and time t (a specific month and a year); predict the most likely event of [Rebellion, Insurgency, Domestic Crisis, Ethnic or Religious Violence, International Crisis] at time t:

$$\text{Crisis}_C(t) = softmax(W * Cr), \qquad \boldsymbol{Cr} = \begin{bmatrix} \text{Cr}_{1_C}(t) \\ \text{Cr}_{2_C}(t) \\ \text{Cr}_{3_C}(t) \\ \text{Cr}_{4_C}(t) \\ \text{Cr}_{5_C}(t) \end{bmatrix} = ?$$

The Second Part: **Prediction Of Major Events**

In the second part, we want to predict the probability of larger subset of major events instead of a small subset of all kinds of crises.
We will simplify the analysis by focusing on a small subset of event sequence candidates that may be causally linked and define sets of events ev_i that are linked to target events ev_j in this manner.

The Problem:

Given some future event ev_j at time τ + Δ and past event ev_i happening at time τ (e.g., today); predict the probability of event ev_j:

$$P\big(ev_j(\tau + \Delta) \,\big|\, ev_i(\tau)\big) = ?$$

Next, we will expand this problem:

In the last problem, we defined a series of events as linked if these are based on documents with similar text and simplified the analysis by focusing on those small subsets of events.

Now, we will define a series of events as linked if those events simply occur or involve a given country. This way, we will have a more vast and rich subset of events (but unfortunately more unrelated events in the subset of each country)

The Problem:

Given some future event ev_{j_C} at time τ + Δ and past events $\{ev_{i_{n_C}}\}_{n=1}^{N}$ happening at time τ, ..., τ-k (a given time window) in a country C; predict the probability of event ev_{j_C}:

$$P\big(ev_{j_C}(\tau + \Delta) \,\big|\, ev_{i_{1_C}}(\tau), \dots, ev_{i_{N_C}}(\tau - k)\big) = ?$$

Datasets:

1. **GDELT**

Source: https://www.gdeltproject.org/

2. **ICWES** (EOI Dataset)

Sub: ICWES (Event Dataset)

Source: https://dataverse.harvard.edu/dataverse/icews

3**. NYT Archive**

Solutions:

We will expend (if needed) the solutions presented in the papers:

For the first part - http://people.cs.vt.edu/naren/papers/websci-gdelt-2014.pdf

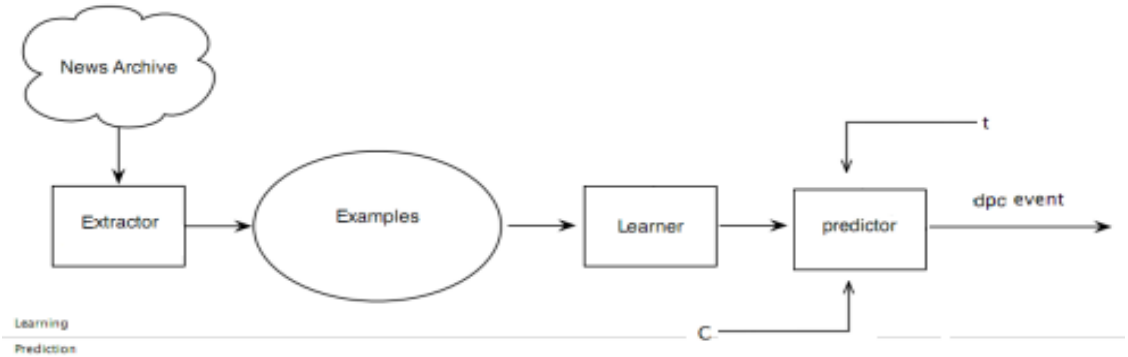
For the second part - http://www.kiraradinsky.com/files/Radinsky-webtorealworld.pdf

The solution for the first part is related to a Graph-Based approach – a graph of all the actors (the countries) and the interactions between them.

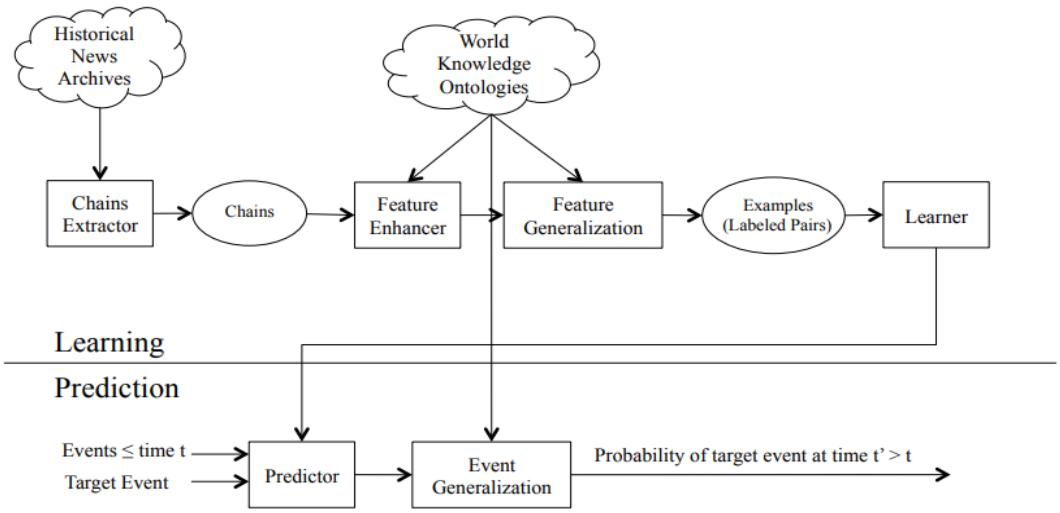
The solution for the second part is related to representing events with lexical and factual features and use Naïve Base approach for the prediction.

The Solution Architectures:

Solution Architecture (The First Part):



Solution Architecture (The Second Part):



Additional Solutions:

1. We will use the interactions between pairs or triples of actor types in each country.

2. Investigate new ways to find related events and create subsets we can work with.

3. Deep Learning and Memory Models such as LSTM and GRU.

Experiments and Evaluations:

The testing methods are similar for each of the 2 problems presented, as we evaluate the predictor's success rate after training it on the old couple of decades and testing it on the most recent years, keeping in mind that there's probably an imbalance between the number of DPC and non-DPC months, and major events and insignificant events.

Framework:

We will be using Python Programming Language coupled with sklearn, TensorFlow and their respective libraries, offering us an extensive machine learning, deep learning and data mining methods for our use.