Non-neural Structured Prediction for Event Detection from News in Indian Languages

Shubhanshu Mishra https://shubhanshu.com
Team: 3Idiots
Code: https://github.com/socialmediaie/EDNIL2020
Presentation at Event Detection from News in Indian Languages (EDNIL) at FIRE 2020 (20st Dec, 2020)

Shubhanshu Mishra, Non-neural Structured Prediction for Event Detection from News in Indian Languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020

Overview

- Participated in both tasks
- Model both tasks as a **single** Sequence Tagging task
- Convert the data into the right format
- Model: N-gram + Regex features → CRF
- Post-process single model predictions for each task submission
- Best performance across all languages on both tasks
- Beats neural models by a huge margin
- Super easy to implement, very fast to train and perform inference
- Open sourced code: https://github.com/socialmediaie/EDNIL2020

Task Overview

Source: https://ednilfire.github.io/ednil/2020/index.html

The goal of this track is to detect events from news articles written in **Hindi, Bengali, Marathi, Tamil** and **English**.

Definition of Event: An event can be an occurrence happening in a certain place during a particular interval of time with or without the participation of human agents. It may be part of a chain of occurrences or an outcome or effect of preceding occurrence or a cause of succeeding occurrences. An event can occur naturally or it can be because of human actions.

An event can have a location, time, agents involved (causing agent and on which the effect of the event is felt) etc. The main event can have sub-events as well.

Task 1: Event Identification: Identify a piece of text from news articles that contain an event.
Task 2: Event Frame: Create an event frame from the news article containing the following details:

1. **Type**: Type and subtype of the line containing the event
2. **Casualties**: No of people is injured or killed/Damages to properties
3. **Time**: When the event takes place
4. **Place**: Where the event takes place
5. **Reason**: Why and how the event takes place

Task 1 Example

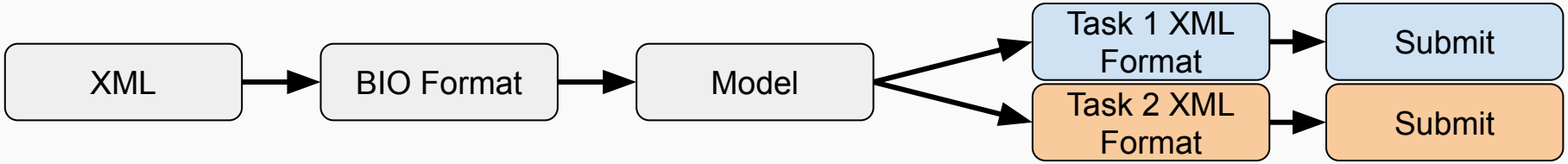
1 dead, 18 hurt in explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at natural gas plant An explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT on Tuesday at a natural gas facility near Austria's border with Slovakia left one person dead, authorities said. A further 18 people were injured in the morning blast MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at the plant in Baumgarten an der March, east of Vienna, regional Red Cross official Sonja Kellner said. Two medical helicopters were sent to the scene, the Austria Press Agency reported. The explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT set off a fire, which operator Gas Connect said was contained by midmorning. The facility was shut down, Gas Connect spokesman Armin Teichert said. Police wrote on Twitter that the situation "is under control." There was no immediate word on what caused the blast MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at the plant, where pipelines connect and gas from Russia, Norway and other countries is compressed.

Task 2 Example

1 dead, 18 hurt CASUALTIES-ARG in explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at natural gas plant PLACE-ARG An explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT on Tuesday TIME-ARG at a natural gas facility near Austria's border with Slovakia PLACE-ARG left one person dead, CASUALTIES-ARG authorities said. A further 18 people were injured CASUALTIES-ARG in the morning TIME-ARG blast MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at the plant in Baumgarten an der March, east of Vienna, PLACE-ARG regional Red Cross official Sonja Kellner said. Two medical helicopters were sent to the scene, the Austria Press Agency reported. The explosion MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT set off a fire, AFTER_EFFECTS-ARG which operator Gas Connect said was contained by midmorning. The facility was shut down, Gas Connect spokesman Armin Teichert said. Police wrote on Twitter that the situation "is under control." There was no immediate word on what caused the blast MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at the plant, PLACE-ARG where pipelines connect and gas from Russia, Norway and other countries is compressed.

Approach

- Single model for both tasks
- Same approach for all languages.




```
for lang in ["en", "bn", "tm", "ma", "hn"]:
    print(lang)
    df_train = pd.read_json(f"./data/processed/{lang}/train.json", orient="records", lines=True)
    X_train, y_train = df_to_Xy(df_train)
    print(df_train.shape)
    crf = sklearn_crfsuite.CRF(algorithm='ap', max_iterations=100, all_possible_transitions=False)
    print("Training")
    %time crf.fit(X_train, y_train)
    df_test = pd.read_json(f"./data/processed/{lang}/test.json", orient="records", lines=True)
    print("Writing")
    create_files(df_test, lang)

```

Code: https://github.com/socialmediaie/EDNIL2020/blob/master/Train%20models.ipynb

Pre-processing and Post-processing

```
1/B-CASUALTIES-ARG dead,/I-CASUALTIES-ARG 18/I-CASUALTIES-ARG hurt/I-CASUALTIES-ARG in/O
explosion/B-MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT at/O natural/B-PLACE-ARG gas/I-PLACE-ARG plant/I-PLACE-ARG An/O
explosion/B-MAN_MADE_EVENT.INDUSTRIAL_ACCIDENT on/O Tuesday/B-TIME-ARG at/O a/O natural/B-PLACE-ARG gas/I-PLACE-ARG
facility/I-PLACE-ARG near/I-PLACE-ARG Austria's/I-PLACE-ARG border/I-PLACE-ARG with/I-PLACE-ARG Slovakia/I-PLACE-ARG
left/O one/B-CASUALTIES-ARG person/I-CASUALTIES-ARG dead,/I-CASUALTIES-ARG authorities/O said./O
```

Preprocessing

- Convert original XML data to BIO format for sequence tagging training.

Post-processing

- For sub-task 1 we only restrict the labels to MAN_MADE_EVENT and NATURAL_EVENT. This allows for using the same structured prediction formulation for prediction for this sub-task.
- For sub-task 2 we use all the labels as above while stripping out the ARG parts from the BIO labels.
- All labels were converted to uppercase to make the labels consistent as few XML labels were lower-cased

Model

Features

- Lower-cased token
- 2 and 3 char suffixes
- If the token is upper cased
- If the token is title cased
- If the token is a digit
- If the token is the beginning of sentence or end of sentence.
- Same features as above for the previous and next tokens.

Model

- Conditional Random Field model (state of the art approach for sequence tagging)
- Trained using averaged perceptron algorithm
- Training time is around 4 minutes per language.

Results

Lang	Task	Team	Precision	Recall	F1 Score
English	1	Ours	0.793	0.703	0.745
English	1	Other best	0.611	0.645	0.628
English	2	Ours	0.504	0.447	0.474
English	2	Other best	0.201	0.248	0.222
Bengali	1	Ours	0.705	0.553	0.620
Bengali	1	Other best	0.379	0.391	0.385
Bengali	2	Ours	0.548	0.411	0.469
Bengali	2	Other best			
Hindi	1	Ours	0.685	0.569	0.622
Hindi	1	Other best	0.505	0.517	0.511
Hindi	2	Ours	0.472	0.341	0.396
Hindi	2	Other best			

Lang	Task	Team	Precision	Recall	F1 Score
Tamil	1	Ours	0.692	0.676	0.684
Tamil	1	Other best	0.138	0.228	0.172
Tamil	2	Ours	0.506	0.469	0.487
Tamil	2	Other best			
Marathi	1	Ours	0.609	0.434	0.507
Marathi	1	Other best	0.124	0.417	0.191
Marathi	2	Ours	0.387	0.278	0.324
Marathi	2	Other best			

Data Statistics

lang	bn	en	hn	ma	tm
test	204	206	160	265	257
train	800	828	677	1030	1013

- Our approach performed **best across all languages and across all tasks**.
- The margin compared to the next best team was often >15% F1-score.
- Task 1 easier than task 2, because labels of task 1 are subset of task 2.
- Marathi with the largest dataset has worst F1 score, but Tamil (2nd largest) is best on task 2.
- Model may have an English bias because of regex features.

Closing thoughts

- **Simple** and **fast baselines** should be tried before implementing more complicated models.
- A single model performs both tasks (across languages) by combining it with post-processing of the model output. This is a computationally cheap way to perform **multi-task inference**.
- Other teams seem to have used neural network models but this simple solution sets a strong baseline others can aim to improve upon.
- A more principled approach can be investigated for integrating neural models for these tasks which has small datasets.

Thank You / Questions

Contact: Shubhanshu Mishra (@TheShubhanshu)

Code: https://github.com/socialmediaie/EDNIL2020

Citation: Shubhanshu Mishra, Non-neural Structured Prediction for Event Detection from News in Indian Languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020

Other tools for information extraction: SocialMediaIE - Social Media Information Extraction | Tools for efficient social media information extraction using advanced machine learning techniques - https://socialmediaie.github.io/