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

Shubhanshu Mishra[a]

[a] *https://shubhanshu.com*

Abstract

In this paper, I give a description of my submission, with team name 3Idiots, to the Event Detection from News in Indian Languages (EDNIL) 2020 shared task. My method is based on structured prediction using only word n-gram and regex features and does not rely on any latest deep learning or neural network methods. The methods was used for both tasks across all languages. It was the best performing system on all tasks and languages, outperforming other best methods by a clear $\tilde{1}0$ to 15 F1-score points. The approach presented here is quite fast to train and do inference. The code will be open sourced at https://github.com/socialmediaie/EDNIL2020.

Keywords

News, Indian Languages, Structured Prediction, Information Extraction, Event Detection, CEUR-WS

1. Introduction

The goal of this paper is to describe my submission under team name 3Idiots to the Event Detection from News in Indian Languages (EDNIL) 2020 shared task [1]. The approach presented here utilizes the structured prediction formulation of natural language processing tasks[2]. Structured prediction problems can be solved using Conditional Random Fields (CRF) [3] which is a popular approach used for structured prediction tasks. The CRF implementation is based on the CRFSuite library [4] was utilized via the Sklearn-CRFSuite library[1]. For features to the CRF model, I only relied on word n-gram and simple regex based features. Since, the method has no language specific feature, it was used for all languages and was extended to both tasks. This approach is inspired from an earlier work on identifying named entities in social media text which also utilized CRF for the prediction [5, 6]. The code will be open sourced at https://github.com/socialmediaie/EDNIL2020.

2. EDNIL 2020 tasks

The Event Detection from News in Indian Languages (EDNIL) 2020 shared task consisted of two sub-tasks.

Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad

mishra@shubhanshu.com (S. Mishra)

https://shubhanshu.com/ (S. Mishra)

0000-0003-0552-8659 (S. Mishra)

[1] https://github.com/TeamHG-Memex/sklearn-crfsuite

split	Bengali	English	Hindi	Marathi	Tamil
train	800	828	677	1030	1013
test	204	206	160	265	257

Table 1
Number of documents for each language and data split

1. **Event Identification**: Identify the piece of text from news articles that contain an event. The events are of two type: Manmade Disaster and Natural Disaster
2. **Event Frame**: From a news article extract the words associated with the following parameters:
   a) Type: Type and subtype of the line containing the event
   b) Casualties: No of people is injured or killed/Damages to properties
   c) Time: When the event takes place
   d) Place: Where the event takes place
   e) Reason: Why and how the event takes place

The EDNIL 2020 task was conducted for five languages i.e. Hindi, Bengali, Marathi, Tamil and English. The distribution of the EDNIL 2020 data is shown in table 1.

3. Method

Both the tasks can be considered a case of structured prediction or information extraction from natural language text [2, 7, 6]. More specifically, this task can be converted to a sequence prediction task where for each word in the sentence we assign a label which identifies which label type does the word belong to. The labels follow the BIO format described in the next section. This format of labels allows us to efficiently extract contiguous token sequences as phrases tagged with labels. I only submitted a single run for each task and language.

3.1. Pre-processing

First the original text documents were converted to a sequence labeling format. The sequence labeling format converts a span of text labeled with **LABEL** to a sequence of tokens each labeled with either **B-LABEL** if the token is the first token of the text span and **I-LABEL** if it any other token of the text span. For text which is not labeled all its tokens are labeled as **O**. This labeling format is known as the BIO format. An example of a text which is represented as tokens, after converting to BIO format is shown below in (format is token/tag) in figure 1.

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 upper-case to make the labels consistent as few XML labels were lower-cased.

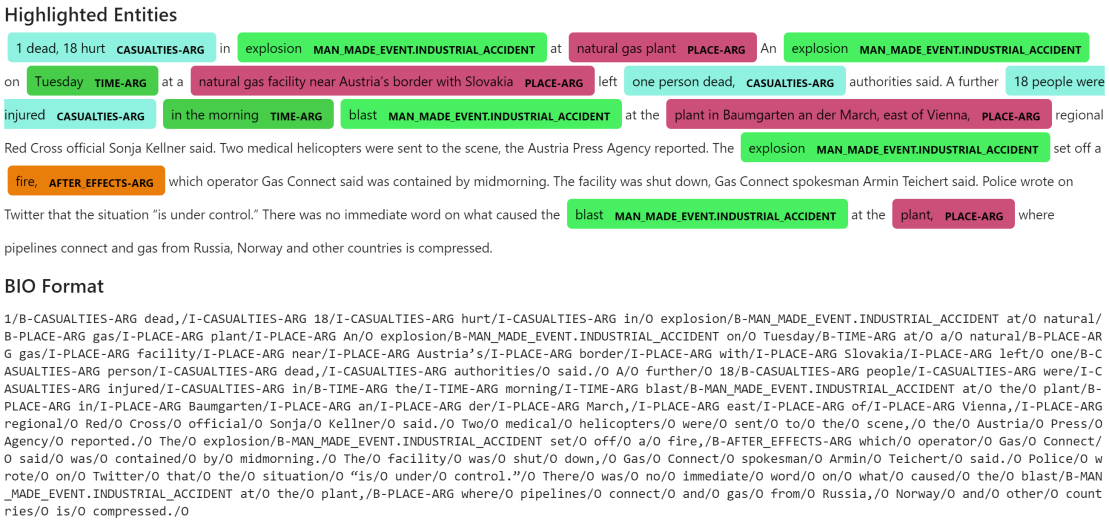

Figure 1: Example of text with highlighted entities and its representation in BIO format for training the model.

3.2. Model

Our model took as input a pair of token features and their labels. Each token in a document was converted to the following 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.

The model was a CRF model was trained using the averaged-perceptron algorithm for a maximum of 100 iterations. A single model was trained for a given language whose results were post-processed for each sub-task. This single model for multiple tasks is inspired from our earlier work on using multi-task models for social media information extraction [5, 6, 7, 8, 9, 10]. This work is also inspired from using simpler approaches for information extraction tasks [11, 12] in this age of neural network models, while continuing to get robust and high quality performance. For each language, the training time was around 4 minutes, which is quite fast compared to many neural network based approaches.

4. Post-processing

The model predictions were post-processed given the task. For task 1, only the text spans which belong to the task specific labels **MAN_MADE_EVENT** and **NATURAL_EVENT** were con-

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			
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			

Table 2
Test data results of our model compared to the other best model (refer to Dave et al. [1] for details) on each language and task.

sidered. These labels were mapped to **MANMADE_DISASTER** and **NATURAL_DISASTER** to make them compatible with the submission format.

5. Results

The submission evaluation on the test data is shown in table 2. Our approach achieved the top spot across all languages and sub-tasks outperforming other solutions by a clear 10-15 % points. Our best performing model for task 1 is English with F1 score of 0.745 which is 12% points higher than the second best submission. The worst performing task 1 model is for Marathi with F1 score of 0.507 which is still 30% points higher than the other best solution. Similarly, for task 2, the top performing model is Tamil with F1 score of 0.487, while the worst performing model is again Marathi with F1 score of 0.324. The performance is always higher on task 1 compared to task 2, because task 2 is a super-set of task 1 and has larger number of labels. One reason for consistent high performance of English model might be the usage of upper-case features which are only applicable for English language. One surprising feature of the model performance is that languages with largest training data like Marathi didn't have the top performance compared

to English which has less data, and Hindi which has the least amount of training data.

6. Conclusion

I have described the approach of the team 3Idiots for the EDNIL 2020 shared task. The approach described achieved the top spot across all languages and tasks using a simple modeling approach. The method is generic and can be improved by using more advanced features for getting token features using the latest deep learning models similar to the approaches taken in [5, 6, 7]. The highlight of the existing method is that a single model can be used via post-processing for both the tasks, which makes the approach quite efficient. This idea suggests an information extraction paradigm which focuses on solving a more complex task and then simplifying it by post-processing of labels. The work presented here is effective yet invites scope for further improvement as well as error analysis. In particular it might be useful to assess the model for inherent biases towards certain types of events similar to the work on bias assessment of named entity recognition systems for demographic biases as shown in [8]. The code will be open sourced at https://github.com/socialmediaie/EDNIL2020.

References

[1] B. Dave, S. Gangopadhyay, P. Majumder, P. Bhattacharya, S. Sarkar, S. L. Devi, Overview of the FIRE 2020 EDNIL track: 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.

[2] S. Sarawagi, Information Extraction, Foundations and Trends® in Databases 1 (2007) 261–377. URL: http://www.nowpublishers.com/article/Details/DBS-003. doi:10.1561/1900000003.

[3] J. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, volume 8, Williams College, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=655813. doi:10.1038/nprot.2006.61. arXiv:arXiv:1011.4088v1.

[4] N. Okazaki, CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. URL: http://www.chokkan.org/software/crfsuite/.

[5] S. Mishra, J. Diesner, Semi-supervised Named Entity Recognition in noisy-text, in: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 203–212. URL: http://aclweb.org/anthology/W16-3927.

[6] S. Mishra, Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets, in: Proceedings of the 30th ACM Conference on Hypertext and Social Media - HT '19, ACM Press, New York, New York, USA, 2019, pp. 283–284. URL: http://dl.acm.org/citation.cfm?doid=3342220.3344929. doi:10.1145/3342220.3344929.

[7] S. Mishra, Information Extraction from Digital Social Trace Data with Applications to

Social Media and Scholarly Communication Data, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2020. URL: http://hdl.handle.net/2142/107965.

[8] S. Mishra, Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data, ACM SIGIR Forum 54 (2020). URL: http://sigir.org/wp-content/uploads/2020/06/p19.pdf.

[9] S. Mishra, S. Prasad, S. Mishra, Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 120—-125. URL: https://www.aclweb.org/anthology/2020.trac-1.19.

[10] S. Mishra, S. Mishra, 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages, in: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, Kolkata, India, 2019, pp. 208–213. URL: http://ceur-ws.org/Vol-2517/T3-4.pdf.

[11] S. Mishra, S. Mishra, Scubed at 3C task A - A simple baseline for citation context purpose classification, in: Proceedings of the 8th International Workshop on Mining Scientific Publications, Association for Computational Linguistics, Wuhan, China, 2020, pp. 59–64. URL: https://www.aclweb.org/anthology/2020.wosp-1.9.

[12] S. Mishra, S. Mishra, Scubed at 3C task B - A simple baseline for citation context influence classification, in: Proceedings of the 8th International Workshop on Mining Scientific Publications, Association for Computational Linguistics, Wuhan, China, 2020, pp. 65–70. URL: https://www.aclweb.org/anthology/2020.wosp-1.10.