# Event Extraction from social media Text in Malayalam using Neural Conditional Random Fields

[1] Pattabhi RK Rao, [2] Sobha Lalitha Devi
[1,2]AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
[1] pattabhi@au-kbc.org, [2] sobha@au-kbc.org

## ABSTRACT

This paper describes a Neural Conditional Random Fields (NCRF) approach for Event extraction (EE) task which aims to discover different types of events along with the event arguments from the user generated text content (tweets) in Malayalam. The data for this work was obtained from FIRE (Forum for Information Retrieval and Evaluation) 2017 shared task [12] on Event Extraction from Newswires and Social Media Text in Indian Languages. A NCRF is a combination of Recurrent Neural Network (RNN) and Conditional Random Fields (CRF). In addition to event detection, the system also extracts the event arguments which contain the information related to the events such as when (Time), where (Place), Reason, Casualty, After-effect etc). Our proposed Event Extraction system achieves F-score of 79.74%. The results are encouraging and comparable with the state-of-art.
Keywords: Event Extraction, Social Media Text, Indian Languages, Malayalam, Neural Conditional Random Fields (NCRF)

## INTRODUCTION

In India and across the world with advancement of technology and hardware becoming cheaper, access to mobile devices such as smart phones, tabs etc., and internet has significantly changed the way people communicates. Facebook and Twitter are two most popular social media platforms, where people post about events, their personal activities, opinions, and ideas. And also post their thoughts, responses or reactions for any public cause or issue. Thus, it is very important to identify relevant events and extract the temporal aspects about those events. Understanding events and their descriptions in raw text is the key factor in automatic event extraction, and is important and challenging task in Natural Language Processing (NLP). Event extraction aims to detect, from the text, the occurrence of events of specific types, and to discover the arguments (event participants or attributes) that are associated with the event. Event Arguments represent the event related information i.e. capturing who does what to whom, how, when and where. It is also essential in practical applications like news summarization, information retrieval and knowledge base construction. Event extraction has been actively researched for over last decade. Most of the research has, however, been focused on resource rich languages, such as English, French and Spanish. The scope of this work covers the task of event recognition and extraction in newswire, social media text such as facebook for Indian languages using the data provided by EventXtract-IL 2017 FIRE 2017 evaluation track[12].

It is observed that there are very little works in Indian language event extraction. The user generated texts such as twitter and facebook texts are diverse and noisy. These texts contain non-standard spellings and abbreviations, unreliable punctuation styles. Apart from these writing style and language challenges, another challenge is concept drift [3][5] the distributionof language and topics on Twitter and Facebook is constantly shifting, thus leading to performance degradation of NLP tools over time.

Further the paper is structured as follows, in section 2, a brief overview of the recent published work is given and section 3 details the features and the methods used in our approach for event extraction. The evaluation and results are discussed in section 4. The paper ends with the conclusion

## RELATED WORK

In recent years Deep Learning is widely used ML methodology for NLP applications. By using the multilayer neural architecture it can learn the hidden patterns from the enormous amount of data and handles the complex problems. Deep learning networks can be roughly categorized into (1) unsupervised/generative, e.g., restricted Boltzmann machines (RBMs) [13], deep belief networks (DBNs)[6]; (2) supervised/discriminative, e.g., deep neural networks (DNNs)[9], convolutional neural networks (CNNs)[7] and recurrent neural networks(RNNs)[14]; and (3) hybrid, e.g., DBNDNN[4] models that combine unsupervised pre-training and supervised fine-tuning.

Learning methods used in event extraction falls into three categories: (1) learning from labeled data (i.e. supervised learning); (2) learning from unlabeled data (i.e. semi-supervised and unsupervised learning); (3) Hybrid approach where learning scheme integration to integrate different learning paradigms at outer system level. Several approaches used in event extraction are Conditional random fields (CRF) [17] and support vector machines (SVM) [15] which are supervised learning methods, and deep neural networks [2] which are unsupervised approach and these have been applied to both general domain information extraction and domain specific such as biology, biomedical etc. One of the open source platforms available to do information extraction for English language documents is the Open Information Extraction (OpenIE) [1], which has emerged as a novel information extraction paradigm. The OpenIE system consist of four main components: (1) Automatic Labeling of data using heuristics or distant supervision; (2) Extractor Learning using relation-independent features on noisy self-labeled data; (3) Tuple Extraction on a large amount of text by the Extractor; (4) Accuracy Assessing by assigning each tuple a probability or confidence score.

## OUR APPROACH

The proposed event extraction system has three components 1) entity identification (NEs) 2) event extraction and 3) event argument extraction. The system follows a pipeline architecture, where the data is first pre-processed to the required format that is needed to train the system. After training the system the NEs are automatically identified from the test set. The following section gives in detail the pre-processing that needs to be done.

### A. Pre-processing

The data, input to the system, is pre-processed for formatting, where we use a sentence splitter and tokenizer and also it is converted into column format. The formatted data is further annotated for syntactic information which includes the Part-of-speech (POS) and Phrase Chunk (Noun Phrase, Verb phrase) tagging. The POS tagger and Chunker for Malayalam were developed in-house using hybrid approach, as described in the work of [11].

### B. Entity Identification

We identify the entities using Neural Conditional Random Fields (NCRFs). Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting sequential data, based on the conditional approach. Lafferty et al. [8] had first used CRFs for NLP applications. A CRF is a form of undirected graphical model or Markov random field, globally conditioned on X that defines a single log-linear distribution over label sequences given a particular observation sequence.

Neural CRFs (NCRFs) is designed with three layers: a character sequence layer; a word sequence layer and inference layer. For each input word sequence, words are represented with word embeddings. The character sequence layer can be used to automatically extract word level features by encoding the character sequence within the word. In this we can also incorporate hand crafted features such as capitalization, suffixes etc.

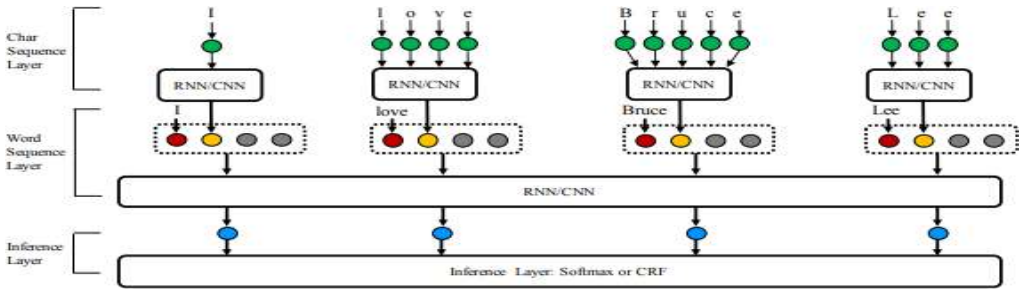


Figure -2. NCRF architecture for an example sentence. Green, red, yellow and blue circles represent character embedding's, word embedding's, character sequence representations and word sequence representations, respectively. The grey circles represent the embedding's of sparse feature

Feature selection plays an important role in the performance of any machine learning system. Also, the features selected must be informative and relevant. We have used word, grammatical and functional level terms as features and they are detailed below:

**Word level features**: Word level features include Orthographical features and Morphological features.

**a. Orthographical features:** contain capitalization, foreign words, combination of digits,symbols.

**b. Morphological/suffix features:** Morphological suffixes of nouns and verbs for example"kal", "il", "ed" etc.

**Grammatical features**: Grammatical features include words, POS, chunks and combination

of words, POS and chunk.

The NCRF++ tool is used for implementation. It is an open source implementation of NCRFs [16]and is a general purpose tool. The features required for training have been explained above in this section. It learns the patterns of named entities from the tagged corpus and using the model generated using the training data the NEs in the test data can be automatically identified. All the features used are extracted from the training corpus provided by the EventXtract-IL 2017 FIRE Track and no other external resources have been used.

### C. Event and Event Argument Extraction

The event and its arguments are extracted for each event. An event can have associated sub-events, but here we do not link events and sub-events. A sub-event is also considered as an event. We have used NCRFs for extracting events and their arguments. The arguments of events give us information about the doer of the event, where and when the event has happened. The main challenges in the event argument extraction are:

i) Capturing the long range connection between the event trigger and event argument

ii) Identifying the correct role of the event argument with respect to the event type (or theevent trigger), and

iii) The span of the argument.

The grammatical features of POS and Named Entities are used for the identification of Events. The NEs identified in the previous step form the arguments of the event. The motivation behind using the word, POS and NE tags is that it can detect the structures in the input and automatically obtain better feature vectors for classification. Most of the earlier NLP works have used words as input for training.

The POS and NE tags help to add sense and semantic information to the learning. The NE tag will help in identifying whether they are attributes of objects, phenomenon's, events etc. This gives clue of the event triggers, while learning and thus help in the identification of the events. We have modeled NCRF as pairs of 3-ary observations. The 3-ary consists of word, POS and NE (Entity Tag).

These three levels of data in the visible layer (or input layer) are converted to vectors of n-dimension and passed to word sequence layer of NCRF. The word vectors, POS vectors and NE vectors are the vector representations. These are obtained from the word2vec. We make use of the DL4J Word2vec API for this purpose [10].

## EVALUATION AND RESULTS

We use the standard evaluation metrics of precision, recall and F measure for evaluating Named Entity recognition and also for Events & their arguments extraction.
The task of extraction of event arguments is modelled as Argument boundary labelling task and the boundary labels used are "Arg1-Start", "Arg1-End", "ArgM-Start" and "ArgM-End". And only when the system identifies all the four boundaries correctly then only the event argument extraction is considered. No partial identification is considered for evaluation.

The training dataset consists of annotated tweets where event trigger words as well as event arguments are tagged. The test set contains of tweets without any annotations. The training file is a column format file, where each column was tab space separated. It consisted of the following columns:

i) Tweet_ID
ii) User_Id
iii) Event string
iv) Event Start_Index
v) EventString_Length

For example:

Tweet_ID:890123456782341
User_Id:987654321
EventString: പരിശീലനത്തിൽ നിർദ്ദിഷ്ടനിലവാരം പുലർത്തുന്നവർക്കു പി ജി ഡിപ്ലോമയും കൂടുതൽ
Index:0
Length:41

The data of Event Extract – IL 2017 FIRE track consisted of 3 Indian languages data. In this work we have only taken Malayalam data. The Malayalam data consisted of 7391 tweets with 1733 events.

Let there be an Event-mention E1 and there are event argument fields such as Event Type, Location, Time, Event-Participants, Causes, and Effects for that event. Now for that event E1, if all these event arguments fields are identified correctly then the system gets full score of 1 else 0. The system obtained a Precision of 78.85% and Recall of 80.65% and F-measure of 79.74%.

## CONCLUSION

In this paper we have described our work on event extraction from social media text in Malayalam. The approach used is scalable and can be used for any language text. The approach uses Neural CRF which combine the power of Neural Networks and CRFs, gives the required flexibility to give features of our own. We have obtained encouraging results which is comparable with state of the art. We have obtained good recall of 80.65 which is significantly better than the results that were submitted in the FIRE 2017 track.

## ACKNOWLEDGEMENTS

The authors would like to thank Event Extract – Indian Languages 2017 FIRE track organizers for giving us the data and helped us conducting the evaluation.

## REFERENCES

[1] Banko M, Cafarella MJ, Soderland S. (2007). *Open information extraction for the web*. IJCAI 2007; 7:2670–2676.

[2] Collobert R, Weston J, Bottou L,. (2011) Natural language processing (almost) from scratch.

*The Journal of Machine Learning Research 2011*; 12:2493–2537

[3] Mark Dredze, Tim Oates, and Christine Piatko, (2010). "We're not in Kansas anymore: detecting domain changes in streams". *In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp 585–595. Association for Computational Linguistics (ACL).

[4] Erhan D, Bengio Y, Courville A. (2010). Why does unsupervised pre-training help deep learning? The Journal of Machine Learning Research 2010; 11:625–660 [6] Dr. Moh. Osama K., "*HELLO Flood Counter Measure for Wireless Sensor Network*," International Journal of Computer Science and Security, vol. 2 issue 3, 2007, pp-57-64.

[5] Hege Fromreide, Dirk Hovy, and Anders Søgaard, (2014). "Crowdsourcing and annotating NER for twitter#drift*". European language resources distribution agency*

[6] Hinton G, Osindero S, Teh Y-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation 2006*; 18:1527–1554

[7] Krizhevsky A, Sutskever I, Hinton GE. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 2012*; 1097–1105

[8] John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proc. 18th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, USA.pp.282-289

[9] Lamblin P, Bengio Y. (2010). Important gains from supervised fine-tuning of deep architectures on large labeled sets. *NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop 2010*

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at ICLR.*

[11] Pattabhi R K Rao T, Vijay Sundar Ram R, Vijayakrishna R and Sobha L. (2007). 'A Text Chunker and Hybrid POS Tagger for Indian Languages'. *In the Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages*, Hyderabad. pp. 9-12.

[12] Pattabhi RK Rao and Sobha Lalitha Devi. (2017). 'EventXtract-IL: Event Extraction from Newswires and Social Media Text in Indian Languages@ FIRE 2017 - An Overview', In the Forum for Information Retrieval and Evaluation-2017.

[13] Salakhutdinov R, Mnih A, Hinton G. (2007). Restricted Boltzmann Machines for Collaborative Filtering. Proceedings of the 24th International Conference on Machine Learning 2007; 791–798

[14] Socher R, Lin CC, Manning C. (2011) Parsing natural scenes and natural language with recursive neural networks. Proceedings of the 28th international conference on machine learning (ICML-11) 2011; 129–136

[15] Tang B, Wu Y, Jiang M. (2013) Recognizing and Encoding Disorder Concepts in Clinical Text using Machine Learning and Vector Space Model. Working Notes for CLEF 2013 Conference 2013; 1179

[16] Jie Yang and Yue Zhang. (2018). NCRF++: An Open-source Neural Sequence Labeling Toolkit. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 74–79 Melbourne, Australia, July 15 - 20, 2018

[17] Uzuner è„°zlem, South BR, Shen S. (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 2011; 18:552–556.