STHAL: Location-mention Identification in Tweets of Indian-context

Kartik Verma[1,*], **Shobhit Sinha**[2,*], **Md Shad Akhtar**[3], **Vikram Goyal**[3]
[1]Delhi Technological University, India
[2]Thapar Institute of Engineering & Technology, India
[3]Indraprastha Instuitute of Information Technology Delhi (IIIT-Delhi), India
vkartik2k@gmail.com, shobhitsinha13@gmail.com
{shad.akhtar, vikram}@iiitd.ac.in

Abstract

We investigate the problem of extracting Indian-locations from a given crowd-sourced textual dataset. The problem of extracting fine-grained Indian-locations has many challenges. One challenge in the task is to collect relevant dataset from the crowd-sourced platforms that contain locations. The second challenge lies in extracting the location entities from the collected data. We provide an in-depth review of the information collection process and our annotation guidelines such that a reliable dataset annotation is guaranteed. We evaluate many recent algorithms and models, including Conditional Random fields (CRF), Bi-LSTM-CNN and BERT (Bidirectional Encoder Representations from Transformers), on our developed dataset named STHAL. The study shows the best F1-score of 72.49% for BERT, followed by Bi-LSTM-CNN and CRF. As a result of our work, we prepare a publicly-available annotated dataset of Indian geolocations that can be used by the research community. Code and dataset are available at https://github.com/vkartik2k/STHAL.

1 Introduction

Location-based systems (Gartner) are playing a vital role in multiple applications such as navigation services, tourist place recommendation, address standardization and safe routes recommendation. To implement such systems and provide location-based services effectively, it requires to have up-to-date information on location names and their associated events. One relevant source for such information is social media which is considered as a crowd-sourced dataset. Social media has become the most potent medium for the real-time source of data (B. Han and Baldwin, 2014) for analysis.

The impact of social media is inevitable and massive. Many case studies reflect on why people are such a lot active on social platforms and share information. This puts the means to connect anywhere, at any time.[1] Journalists also use it as a medium of power to gain insights about the background of various events. However, social networks enable one to harvest recent location events; there lie various challenges due to the platforms being general in terms of sharing of information by individuals. One of the challenges is to select relevant posts out of streaming data that contain location information. The second challenge is to extract location entities from noisy text (Kumar and Singh, 2019) data.

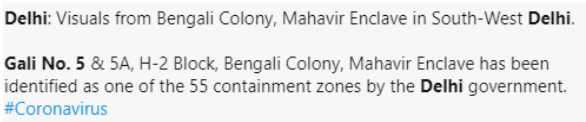

Figure 1: Example of a tweet having non-standard location name

Previous state-of-the-art techniques based on Geolocation Prediction in Twitter (Chi et al., 2016), Detecting Location-Centric Communities (Lim et al., 2015) and Social Media Data Location Prediction (Han et al., 2012) do not perform well in terms of extracting fine-grained location entities in Indian context from the crowd-source data. One of the specific reasons for inferior performance is the non-standardization in location naming conventions. For example, it is easy to find locations names having words such as *gali*, *zila*, *village*, *vihar*, *nagar*, *gaon*, etc. An example is given in Figure 1. It demands designing of specific methods to extract location names and associated events from crowd-source data for Indian subcontinent. The application of this approach can help to demystify the route system, telematics vehicle tracking, and Covid tracking using crowd sourced data.

* First two authors have contributed equally.

[1] https://www.simplilearn.com/real-impact-social-media-article

Proceedings of the 17th International Conference on Natural Language Processing, pages 379–383
Patna, India, December 18 - 21, 2020.

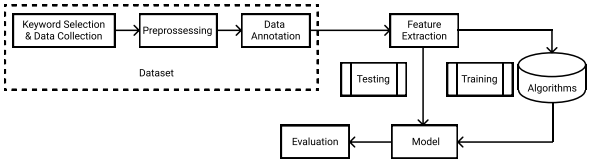

Figure 2: Process flow diagram for the location tagging.

In this paper, we investigate various approaches for named entity recognition focusing mainly on location names. The approaches include Conditional Random Field (CRF) (Lafferty et al., 2001), neural network-based Bi-directional Long Short Term Memory-Convolution Neural Network (Bi-LSTM-CNN) (Chiu and Nichols, 2016), and Bidirectional Encoder Representations from Transformers-based (Devlin et al., 2019) models. We improve the fine-grained location-based approaches by fine-tuning BERT on our annotated dataset (Arase and Tsujii, 2019). We observe a performance improvement of $\geq 2\%$ F1-score points in BERT-based model compared to the other two baseline systems.

The remainder of the paper is organized as follows: In Section 2, we describe the development of STHAL dataset and the proposed approach. We present our experimental results and necessary analyses in Section 3. Finally, we conclude in Section 4.

2 Methodology

In this section, we describe our methodology in detail. First, we explain STHAL's development process, and subsequently, present our system. A high-level diagram is depicted in Figure 2.

2.1 Dataset Development

As discussed above, the need for dataset development is driven by the fact that there exist no[2] tweets dataset to cater the requirements of location-specific entity extraction (or NER, in general) in the Indian context. Therefore, we collect tweet that mentions geographical location (or address) in India and annotates them accordingly.

2.1.1 Data collection

The collection of the dataset was divided into two stages: keyword selection for the seed word list and extraction of tweets for each keyword in the seed list. We adopt the (Aref et al., 2020) method for the collection process.

Keyword Selection In doing the research, the first challenge was to collect comprehensive data for the topic. We mainly focused on collecting the data based on location irrespective of its other attributes, such as the statement's sentiment or the lingual. To tackle this issue, one comprehensive solution can be creating a list of keywords that are concentrated on tweets and crawl using an API.

We created a set of words which was used to extract data from Twitter. We choose *Covid*, *accident*, and *road* as keywords for extraction. These keywords were then iterated with the location database (covered in the next section) to get the required dataset.

Data Set Collection: Among various social media handles, we choose Twitter to extract the dataset because of its wide range of coverage. Twitter is playing an essential role in providing public tweets in the form of JSON document, and it includes various added fields as well such as locations, status Etc.

We use Tweepy[3], a standard Twitter API consist of REST (Representational State Transfer) APIs and Streaming APIs, for data collection. The REST search API provides access to general public tweets with other relevant information. Bulk queries were made to extract the required information. About 250 tweets were only retrieved within the given time frame i.e, 1st January 2020 to 31st August 2020. There was various limitation related to the tweets mentioned in the documentation. Due to these limitations in search API, a custom tweet scraper was made to query the tweets for a given time frame and iterated the keywords and location database (about 233 locations) to get the required database. Around 3500, tweets were retrieved.

2.1.2 Data Pre-processing

Post data collection phase, we apply a series of pre-processing steps to clean our dataset as follows:

- **Removal of irrelevant tweets:** We remove some tweets irrelevant for our case. For example, in the tweet '*Delhi beats Mumbai in Ranji Trophy.*', the mentions of '*Delhi*' and '*Mumbai*' are not referring to a geographical location; instead, they are referring to a cricket team playing for the two cities.

- **Normalization :** In general, tweets consist of lots of noisy texts; therefore, we normalize the

[2]To the best of our knowledge.

[3]https://www.tweepy.org

Text	This	is	the	situation	of	Mahatma	Gandhi	Road	,	Adarsh	Nagar	,	Delhi	33	from	the	last	month	.
Labels	O	O	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	O	O	O	O

Table 1: An example annotated tweet following BIO (Tjong Kim Sang, 2002) scheme from **STHAL dataset**

tweets to remove unprintable, junk, and some special ($, ⋆, Etc.) characters.

Finally, we tokenize the remaining tweets using CMU Ark tokenizer[4] for further processing.

2.1.3 Data Annotation

The annotation process involves the analysis of each tweet in the dataset manual. We adopt BIO notation scheme (Tjong Kim Sang, 2002) to assign a tag (*B-LOC*, *I-LOC*, or *O*) to each token of a tweet. An annotated example is shown in Table 1. The tweet contains three fine-grained locations ('*Mahatama Gandhi Road*', '*Adarsh Nagar*', and '*Delhi 33*') that constitute one coarse-grained location (i.e., '*Mahatama Gandhi Road, Adarsh Nagar, Delhi 33*'). We annotate the location at the fine-grained level. We can observe that the fine-grained locations are separated by punctuation (usually, comma) marks; thus can be easily constructed back to the coarse-grained annotations by assigning *I-LOC* to each intermediate punctuation.

The first token of each location gets a *B-LOC* tags marking the begin of the location. Each subsequent tokens in the location get *I-LOC* tags reflecting the intermediate positions of the location. All non-location tokens are marked with *O* representing outside of the location.

Table 2 lists statistics of the STHAL dataset. In total, it consists of $3,411$ tweets with $8,369$ location mentions.

Stats	**Value**
No. of tweets	$3,411$
No. of tokens	$1,09,162$
No. of locations	8369
Avg. length of sentence	32.002 tokens
Avg. location length	2.255 tokens
Multilingual	English and Romanized Hindi

Table 2: A few statistics of the STHAL dataset.

2.2 Model and Other Baselines

The named-entity-recognition task is a sequence-labelling task, in which, for a sequence of n tokens (i.e., a sentence), we expect a sequence of m tags, where $n == m$. Following the similar setup, we employ BERT (Bidirectional Encoder Representations from Transformer) (Devlin et al., 2019) architecture. To compare the goodness of BERT-based system, we also employ two standard models for sequence labelling task, i.e., a CRF-based model and a Bi-LSTM-CNN (Chiu and Nichols, 2016) architecture.

- **BERT:** BERT as the sequence-learner for the automatic extraction of location mentions from tweets. Recently, BERT has been established as a de facto system for a variety of NLP tasks mainly due to its excellent capability in extracting the underlying semantics from the text. We utilize a pre-trained BERT base model and fine-tune it for the location mention identification in tweets.

- **CRF:** A CRF (Conditional Random Field) (Lafferty et al., 2001) is a class of discriminative model, used for predicting sequences. It exploits the contextual information of the input as well as the predicted labels of the preceding tokens for classifying the current token. The tokens are converted into feature vectors (Quang H Pham, 2019) and are then used by the CRF for sequential labelling. We compute the following three features for the current and previous three tokens: surface form in lower case; a binary feature for all caps; and a binary feature for title case.

- **Bi-LSTM-CNN:** The second system is a pipeline model of Bi-LSTM (Hochreiter and Schmidhuber, 1997) followed by a CNN layer (Kim, 2014). We employ GloVe embeddings (Pennington et al., 2014) model for the feature extraction of input tokens. The hidden representations of Bi-LSTM is fed to a CNN layer and subsequently to the output layer for final classification. To ensure the convoluted features for each token, we zero-padded (Hashemi, 2019) the input. We use 30 trigram filers followed by max-pool (Wu and Gu, 2015) layer.

3 Experiments and Evaluation Results

We implement CRF, Bi-LSTM-CNN, and BERT models in sci-kit-learn, TensorFlow, and PyTorch

[4] http://www.cs.cmu.edu/~ark/TweetNLP/

	Text	Check	distance	from	Gali	No	.	5	,	Dwarka	to	Subzi	Mandi	Old	,	New	Delhi	,	Delhi	-	110036	.
	Gold	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*O*	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O
	CRF	O	O	O	O	O	O	O	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	O	O	O	O
A	**B-CNN**	O	O	O	*B-LOC*	O	O	O	O	O	O	O	O	O	O	*B-LOC*	*I-LOC*	O	*B-LOC*	O	O	O
	BERT	O	O	O	*B-LOC*	*I-LOC*	O	O	O	*B-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	*B-LOC*	O	*I-LOC*	O
	CRF	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	B-LOC	O	O	O
B	**B-CNN**	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O
	BERT	O	O	O	*B-LOC*	*I-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	O	*B-LOC*	*I-LOC*	*I-LOC*	O

Table 3: A qualitative analysis of the obtained outputs for two setups, A and B. We make two observations: a) Training on dataset with Indian addresses and locations help; and b) BERT yields better outputs (it correctly identifies all five instances of location mentions in setup B) compared to the other two baselines. Red text marks misclassifications.

libraries, respectively. For the evaluation, we utilize CONLL-2002 evaluation script[5] for computing the precision, recall, and F1-score for the location mentions. In all the experiments, we randomly split our annotated dataset, STHAL, into 75:25 ratio for the train and test sets. Moreover, to establish our hypothesis that the exiting NER datasets do not adapt well to the location identifications for the Indian context, we conduct our experiments in two setups.

- **Setup A:** Training on the existing Named Entity Recognition system (NER) [6] dataset and testing on the STHAL's test set.
- **Setup B:** Train and testing on STHAL.

In Table 4, we report our experimental results for both setups on STHAL's test set. All three models, i.e., CRF, Bi-LSTM-CNN, and BERT, yield F1-scores of 19.60%, 26.08%, and 34.03%, respectively, in setup A. One important observation we make here is that the precision of CRF is the highest, while recall is the lowest among all. This suggests that the CRF model is too pessimistic about tagging a token as location-mention, i.e., the low recall value reflects the non-aggressive approach in tagging tokens as location-mentions, and the tokens it tagged as location-mentions are correct 66.23% (precision) of times. The BERT-based model improves upon the recall value but at the cost of low precision; however, the F1-score also improves.

It is evident that the best model in setup A (BERT) does not have good F1-score, mainly due to lack of Indian-styled location-mentions in the train set. In comparison, we observe significant improvements for all models in setup B. The best F1-score of 72.49% is obtained by BERT, followed by Bi-LSTM-CNN (70.31%) and CRF (69.99%).

[5] https://www.clips.uantwerpen.be/conll2002/ner/bin/conlleval.txt

[6] https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus

Setup	Model	Precision	Recall	F1-Score
A	CRF	66.23%	11.50%	19.60%
	Bi-LSTM-CNN	41.26%	19.06%	26.08%
	BERT	34.03%	27.36%	34.03%
B	CRF	75.46%	65.26%	69.99%
	Bi-LSTM-CNN	67.45%	74.41%	70.31%
	BERT	71.98%	73.00%	72.49%

Table 4: Experimental results for location-mention identification on STHAL's test set. It's hard-evaluation, i.e., if any token is misclassified in a location mention, we treat it as the misclassified location mention.

3.1 Error Analysis

We present a qualitative analysis of the obtained outputs in Table 3. For an example tweet in STHAL's test set, we list the token-wise prediction for all three systems in two setups. We make the following two observations:

- In setup A, where the systems are trained on the existing NER dataset (covering global addresses and locations), all systems -including BERT- commit mistakes in identifying *Gali No. 5* as location. In contrast, we observe a better performance of these systems when trained on the STHAL dataset (covering Indian addresses and locations) in setup B.
- In both setups, we observe a superior performance of the BERT-based system compared to the other two baseline systems.

4 Conclusion

In this paper, we present our research on location-mention identification in Indian-context. Due to the lack of representation of Indian-styled location-names and addresses (e.g., *Gali No.*, *chowk*, etc.) in existing datasets, we develop a new Twitter dataset, STHAL, for location-mention identification in Indian context. We benchmark STHALusing BERT-based sequence classifier. Evaluation shows that the underlying system leverages the presence of the Indian-styled location-mentions in train set.

In STHAL, we include location-mentions primarily from Delhi-NCR and northern part of India. Thus, we hypothesis that it may not be adequately sufficient for discovering location-mentions across India, e.g., southern or north-eastern India. In future, we would like to explore the task of location-mentions suitable for the entire nation.

References

Yuki Arase and Junichi Tsujii. 2019. Transfer fine-tuning: A bert case study. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.

Abdullah Aref, Rana Mahmoud, Khaled Taha, and Mahmoud Al-Sharif. 2020. Hate speech detection of arabic shorttext. In *9th International Conference on Information Technology Convergence and Services (ITCSE 2020)*, pages 81–94.

P. Cook B. Han and T. Baldwin. 2014. Text-based twitter user geolocation prediction. In *Journal of Artificial Intelligence Research*, volume Vol. 49, No. 1, pages 451–500.

Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J. Butler. 2016. Geolocation prediction in Twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 227–234, Osaka, Japan. The COLING 2016 Organizing Committee.

P.C. Chiu, Jason and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Krisp Jukka M. Raubal Martin Van de Weghe Nico Gartner, Georg. Location based services: ongoing evolution and research agenda.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.

Mahdi Hashemi. 2019. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International Journal of Disaster Risk Reduction*, 33:365 – 375.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kwan Hui Lim, Jeffrey Chan, Christopher Leckie, and Shanika Karunasekera. 2015. Detecting location-centric communities using social-spatial links with temporal constraints. In *Advances in Information Retrieval*, pages 489–494, Cham. Springer International Publishing.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nguyen Viet Cuong Quang H Pham, Binh T Nguyen. 2019. Punctuation prediction for vietnamese texts using conditional random fields. In *SoICT 2019: Proceedings of the Tenth International Symposium on Information and Communication Technology*, page 322–327, New York NY United States. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Haibing Wu and Xiaodong Gu. 2015. Max-pooling dropout for regularization of convolutional neural networks.