


# DanfeNER: Named Entity Recognition in Nepali Tweets

Nobal Niraula and Jeevan Chapagain




# Outline

- 
- Introduction
  - Related Work
  - Corpus Preparation
  - Methodology
  - Results
  - Conclusion

# Introduction

## Introduction

- 
- Social Media like Twitter and Facebook allow users to express and share opinions in any topics
  - Tweets are useful to share views and opinions on trending topics
  - Analyzing and understanding tweets is important
  - Tweets are rich with named entities like Person, Location, Organization

# Introduction

- Named Entity Recognition (NER) involves identifying & categorizing key entities in a text
- NER is a fundamental task in NLP
- Resource poor language like Nepali, not much study has been done
- NER models developed for formal languages such as News articles do not perform well for tweets (Liu et al. 2011)

#	Original Tweet and Translation
1	पैसो जति एता रान अनि बैंक मा के को पैसा हुनु ? (Money is here, that's why banks don't have money.)
2	लिङ्देन <sup>PER</sup> बाजे नि अब सकिय क्यारे (Lingden is gone too, it seems.)
3	बाँदरको समस्याले वैकल्पिक खेतीतर्फ <b>पाल्पा</b> <sup>LOC</sup> का किसान - <b>kaligandaki Khabar</b> <sup>ORG</sup> (Farmers in Palpa are choosing alternative farming due to monkey problems - Kaligandaki Khabar.)
4	<b>यु १९ महिला विश्वकप</b> <sup>EVNT</sup> छनोट- <b>नेपाल</b> <sup>ORG</sup> ले <b>भोलि</b> <sup>DATE</sup> <b>कतार</b> <sup>ORG</sup> विरुद्ध खेल्ने (Nepal plays against Qatar tomorrow for U-19 Women Cricket World Cup.)
5	Officially break up dherai jasto ko Nepali Cricket prati. (Most people officially break up with Nepali Cricket)
6	Kun geet ho yesto? Dai lai suhayo ta hai! :) :) (Which song is that ? It is a good fit for you brother!)

## Main Contributions



- Benchmark Data Sets
- End-to-end NER model for Nepali Tweets
- Detailed Error Analysis

# Related Work

## Related Work

- Bam and Shahi 2014
  - Used word features as well as gazetteers including person, organization, location, middle name, verb, designation and others
  - Entities covered: Person, Location, Organization
- Dey, Paul, and Purkayastha 2014
  - Used Hidden Markov Model with n-gram technique for extracting POS tags.
  - Combined POS tag, proper noun and common nouns in a gazetteer list as a lookup table



## Related Work

- Singh, Padia, and Joshi 2019
  - Multiple neural models such as BiLSTM, BiLSTMCNN, BiLSTMCRF, and BiLSTMCNNCRF with different word embeddings
  - No annotation guideline and no human evaluation of the annotated corpus i.e. inter-rater agreement
  - Entities Covered: Person, Location, Organization, Misc
- Niraula and Chapagain 2022
  - Detailed guideline to annotate entities and human evaluation of the annotated corpus
  - Benchmark datasets containing separate training and testing set
  - Entities covered: Person, Location, Organization, Event, Date
  - Transformer Model can have state-of-the-art performance in Nepali NER

# Corpus Preparation

## Data Preparation



- Used Tweepy to extract tweets ( query as **'lang:ne'**)
- Twitter API has rate-limit, so we crawled data in multiple independent requests
- Removed HTML tags, https links, hashtags, emojis, mentions
- Filtered tweets out with less than five tokens
- Corpus size: 85,418
- Marked named entities at character level using Label Studio (Tkachenko et al. 2020)

## Data Preparation

देउवाको अमेरिका भ्रमणको विषय बारे जनतालाई किन जानकारी गराइएन?: भीम रावल


Figure 1: Character level annotation for Named Entities using Label Studio

## Annotation Target and Process




- Person (PER), Location (LOC), Organization (ORG), Event (EVT), and Date (DAT)
- Annotation Guidelines provided by Niraula et.al.(2022)
- Inter-rater agreement of 0.75 based on Cohen's Kappa

## DanfeNER Data Sets

- 
- 7,667 annotated tweets in Nepali Language with 4,966 entities in total
  - Used 70-30 split procedure to create train and test data
  - We tokenized text and provided labels per token

## DanfeNER Dataset



Data	No. Tweets	Tokens	Avg. Len	LOC	ORG	PER	EVT	DAT	Total Entities
Train	5,366	92,425	17.22	923	782	1,061	34	663	3,463
Test	2,301	39,133	17.00	389	356	444	28	286	1,503
Total	7,667	131,558	17.11	1,312	1,138	1,505	62	949	4,966

Table 1: DanfeNER Data Set Statistics

# Annotation Guidelines

NE	Guidelines	Examples
PER	Proper names of people including first names, last names, individual or family names, fictional names and unique nicknames. Generational markers such as Jr. and IV are included. <b>DO NOT MARK</b> honorific titles such as titles (डा), relation names (आमा, ममी, मिस ), pronouns (तिमी, उनी ), reflexive pronouns (आफै), name prefixes (श्री, श्रीमान, डा, प्राडा), and royal titles (राजा, रानी, युवराज) and Sir (सर)	(a) First names: e.g. पुष्पकमल, नारायणकाजी (b) Family names: e.g. महारा, शाही (c) Generational markers: जुनियर and पाचौं (d) Aliases, nicknames: e.g. प्रचन्ड, वादल, चरी (e) Combinations of I-4: e.g. नारायणकाजी श्रेष्ठ, पुष्पकमल दाहाल 'प्रचन्ड' (f) Fictional/mythological characters: e.g. रावण, कृष्ण
LOC	All man-made structures and politically defined places like the names of countries, rivers, and railway stations are marked as LOC. <b>DO NOT MARK</b> a generic reference to a location or a nationality e.g. नदी, समुन्द्र, अमेरिकि, नेपाली	(a) Buildings: e.g. पौड्यो घर, एपोलो अस्पताल (b) Cities, towns, city districts: e.g. माईती घर, कोहलपुर, ललितपुर (c) Continents: e.g. एसिया (d) Countries, states: e.g. क्यानडा, प्रदेश ५ (e) Geographical areas: e.g. अन्नपूर्ण क्षेत्र, मेन्ल्यान्ड चाईना (f) Parks: रारा राष्ट्रिय निकुन्ज, गोदावरी (g) Planets, celestial objects: e.g. प्रिथ्वी, चन्द्रमा (h) Seas, lakes, rivers: e.g. वन्नालको खाडि, त्रिसुली
ORG	The name of a company, media group, team, political party or any other entity created by a group of people.	(a) Commercial companies: e.g. नेपाल टेलिकम, गुगल (b) Commissions: e.g. खानेपानी विभाग (c) Communities/groups of people: e.g. लिम्बु सेवा समाज, सांस्कृतिक केन्द्र (d) Education & scientific institutes: e.g. राष्ट्रिय अनुसन्धान केन्द्र, पुल्चोक क्याम्पस (e) Judicial systems: e.g. काठमाडौं जिल्ला अदालत, सर्वोच्च अदालत (f) Law enforcement organizations: e.g. अनुसन्धान विभाग, नेपाली सेना (g) News agencies and stations: e.g. कान्तिपुर दैनिक, हिमालयन टिभी (h) Political parties: e.g. नेपाली कांग्रेस (i) Public administration: e.g. परराष्ट्र मन्त्रालय, युरोपियन युनियन (j) Sport leagues and clubs: e.g. आई सि सि, नेपाल क्रिकेट संघ, रियल मेड्रिड (k) Banks: e.g. सानिमा बैंक (l) Organization websites: e.g. अमेजन डट कम
EVT	Named events and phenomena including natural disasters, hurricanes, revolutions, battles, wars, demonstrations, concerts, sports events, etc.	(a) Expos: e.g. पोखरा कवि गोष्ठी, मोवाईल एक्पो, गैडाकोट महोत्सव (b) Explicitly marked events e.g. टेलिकमको वार्षिक साधारणसभा, चितवन महोत्सव (c) Sporting Leagues e.g. विश्वकप, लालिगा, एफ वान (d) Hurricanes e.g. स्याण्टी हुरिकेन (e) Battles and Revolutions e.g. काँगडा लडाईं, माओवादी युद्ध
DAT	Date or period of 24 hours or more, including day, week, month, certain named period, season, year, etc. Age is also included in this category whether it is a noun, adjective, or adverb phrase. Numerical values can be spelled out or expressed using digits.	(a) Full or partial date: १५ कार्तिक २०७६, असार १५ (b) Duration: हजारौं वर्ष, माघ १२ देखि १५ (c) Age: ३५ वर्षीय, ३५ वर्षका (d) Season: वसन्त रीतु, शिशिर (e) Day and month: आइतवार वैशाख

Table 2 : Annotation guideline for EVERESTNER data set




# Experiments

## Methodology




- Transformers have shown state-of-the-art performance in Nepali NER tasks
- Monolingual Nepali transformer models are trained from scratch using Nepali text while multilingual models are trained to combine other languages
- Five different transformers based models:
  - Npvec1-BERT ( baseline)
  - NepaliBERT
  - NepBERT
  - DB-BERT
  - BERT-bbmu

## Methodology




Notation	Model	Hugging Face Model ID	Tokenizer	Vocab	Train Data	Params
NPVec1-BERT	BERT	nowalab/nepali-bert-npvec1	WP	30,000	Wiki, OSCAR, news	22.5M
NepaliBERT	BERT	Rajan/NepaliBERT	WP	50,000	LSNC, OSCAR	82M
NepBERT	RoBERTa	amitness/nepbert	BBPE	52,000	CC-100	83.5M
DB-BERT	DistilBERT	Sakonii/distilbert-base-nepali	SP	24,581	OSCAR, CC-100, Wiki	67M
BERT-bbmu	mBERT	bert-base-multilingual-uncased	WP	105,879	Wiki, 102 languages	110M

## Experiments

- 
- Model evaluated based on precision, recall and F1-score
  - All models were trained: 10 epochs, learning rate = 0.0001, batch size = 10
  - Baseline system obtained F1-score of 0.63
  - DB-BERT performed best with F1-score of 0.80

# Results


## Model Comparison



Model	Precision	Recall	F1-score
NPVec1-BERT	0.63	0.62	0.63
NepaliBERT	0.72	0.69	0.70
NepBERT	0.71	0.69	0.70
<b>DB-BERT</b>	<b>0.80</b>	<b>0.80</b>	<b>0.80</b>
BERT-bbmu	0.76	0.74	0.75

**Table 3: Models comparison using micro averaged F1-score**


## Best Performing Model Per Named Entities



Entities	Precision	Recall	F1-score	Support
PER	0.81	0.77	0.79	444
LOC	0.83	0.86	0.84	389
ORG	0.79	0.79	0.79	356
EVT	0.53	0.29	0.37	28
DAT	0.78	0.84	0.81	286

**Table 4: Performance evaluation of the best performing model per named entities**

## Applying News NER model on Tweets



	Train Data	Precision	Recall	F1-score
news	EverestNER-Train	0.66	0.76	0.71
tweets	DanfeNER-Train	0.80	0.78	0.79
news & tweet	DanfeNER-Train + EverestNER-Train	0.78	0.83	0.80

Table 5: DB-BERT performance in different training datasets



# Conclusion

## Conclusion

- Systematic study of the Named Entity Recognition problem in Nepali Tweets
- Constructed the DanfeNER data set, the first benchmark data set for building and evaluating NER systems for Nepali ( <https://github.com/nowalab/DanfeNER> )
- Developed the end-to-end NER neural models for Nepali tweets BERT-based architectures
- NER for News does not perform well on Nepali Tweets
- Future work includes: discovering NE in romanized tweets, handling code-switching



Thank You  
for Your Attention

Jeevan Chapagain  
[ichpgain@memphis.edu](mailto:ichpgain@memphis.edu)