


Named Entity Recognition for Nepali: Data Sets and Algorithms

Nobal Niraula and Jeevan Chapagain



Outline

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


Introduction

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

Token	Label
सशस्त्र	B-EVT
प्रहरी	I-EVT
दिवस	I-EVT
मा	O
प्रधानमन्त्री	O
ओली	B-PER
आज	B-DAT
सशस्त्र	B-ORG
प्रहरी	I-ORG
को	O
मुख्यालय	O
हलचोक	B-LOC
मा	O
सम्बोधन	O
गर्दै	O

Main Contributions

- 
- Annotation Guideline
 - Coverage: Five entities
 - Benchmark Data Sets
 - End-to-end NER model

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
- Not clear how the authors generated the training data set
- Context information not taken into account while training the model

- Dey, Paul, and Purkayastha 2014

- Used Hidden Markov Model with n-gram technique for extracting POS tags.
- Does not describe how the training examples are obtained and used
- Combined POS tag, proper noun and common nouns in a gazetteer list as a lookup table

- 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
- Do not provide separate train and test data sets, making it harder for other NER systems to be evaluated and compared against.


Corpus Preparation

Corpus Preparation




- News articles contains frequently used named entities
- 1,000 news articles from setopati.com
- Articles have different domains like: politics, sports, economics, art, society, literature

Data Preparation

- 
- Removed HTML tags
 - Sentences typical end with ‘|’ and other punctuation marks ‘!’, ‘?’
 - Marked named entities at character level using BRAT Annotation Tool (Stenetorp et al. 2012)

Data Preparation



वेदलाई हातमा लिएर नेपाली मूलका **Person** ह्यारीले **Location** अमेरिकामा लिएर शपथ

Figure 1: Character level annotation for Named Entities using BRAT tool

Annotation Target and Process



- Person (PER), Location (LOC), Organization (ORG), Event (EVT), and Date (DAT)
- Created guidelines and annotated based on those guidelines
- Inter-rater agreement of 0.74 based on Cohen's Kappa

EverestNER Data Sets



- 28,281 sentences corresponding to 996 news articles to five annotation target were annotated
- Used 85-15 split procedure using random selection
- Four times more annotated sentences and twice as many entities as the previously available data sets in Nepali NER

EverestNER Dataset

Data	Articles	Sentences	Tokens	Avg. Sen. Len	LOC	ORG	PER	EVT	DAT
Train	847	13,848	268,741	19.40	5,148	4,756	7,707	313	3,394
Test	149	1,950	39,612	20.31	809	715	1,115	59	572
Total	996	15,798	308,353	19.51	5,957	5,471	8,822	371	3,966

Table 1: EVERESTNER 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. DO NOT MARK 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, क्रिष्ण
LOC	All man-made structures and politically defined places like the names of countries, rivers, and railway stations are marked as LOC. DO NOT MARK 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



Baseline Model

- Rule based model
- Makes a lookup dictionary to map an entity token span to its target label
- For prediction, finds the longest token span in input match exactly with a key in the lookup dictionary

BERT-based model

- Created BERT based NER model using NERDA library (Kjeldgaard and Nielsen 2021) called BERT-bbmu
- Uses bert-base-multilingual-uncased (Devlin et al. 2018b), a multilingual BERT model trained with Wikipedia data on 102 languages including Nepali


Methodology




BLSTM-CRF Models

- Configured different BLSTM CRF model architectures using NCRF++ library (Yang and Zhang 2018)
- Generate word features using Word and Character embeddings as well as external rules
- Library can take pre-trained Word and Character embeddings or can itself learn them during training

Results

- 
- Model evaluated based on precision, recall and F1-score
 - Baseline system obtained F1-score of 0.62
 - BLSTM-CRF model trained for 50 epochs with 0.015 learning rate
 - BERT-bbmu model trained for 10 epochs with 0.0001 learning rate
 - High performance by both BERT-bbmu and BLSTM-CRF


Results



Model	Precision	Recall	F1-score
Baseline	0.71	0.55	0.62
BLSTM-CRF-wc.ft	0.89	0.74	0.81
BERT-bbmu	0.87	0.84	0.85

Table 3: Models comparison using micro average scores. Notations: u=Uncased, w=Word, c=Character, ft=fastText

Results



Entities	Precision	Recall	F1-score	Support
PER	0.90	0.85	0.88	1115
LOC	0.85	0.80	0.82	809
ORG	0.85	0.83	0.84	715
EVT	0.46	0.42	0.44	59
DAT	0.91	0.91	0.91	572

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

Conclusion

Conclusion

- First systematic study of the Named Entity Recognition problem in Nepali
- Constructed the EverestNER data set, the first benchmark data set for building and evaluating NER systems for Nepali (<https://github.com/nowalab/everest-ner>)
- Developed the end-to-end NER neural models for Nepali using BLSTM-CRF and BERT-based architectures
- To our knowledge, we are the first to experiment with the BERT-based models for NER in Nepali.

Thank You!!

Jeevan Chappgain

jchppgain@memphis.edu

Experimental Setup



BLSTM-CRF Model:

- 300 dimensional retrained fastText Word Embedding from NPVEC1 (Koirala and Niraula 2021)
- 4 CNN Layers
- Number of epochs: 50
- Learning rate: 0.015
- Batch size: 50
- Dropout: 0.5
- SGD Optimizer

Experimental Setup



BERT-bbmu Model:

- Number of epochs: 10
- Batch size: 10
- Learning rate : 0.0001