Custom Named Entity Recognition for Gujrati Text Using Spacy

Komil B. Vora
Department of Information Technology
V.V.P. Engineering College
Rajkot
komilvora@gmail.com

Dr. Avani R. Vasant
Computer Science & Engineering Department
DRS Kiran & Pallavi Patel Global University (KPGU)
Vadodara
avani.vasant@gmail.com

Dr. Saurabh Shah
Computer Engineering Department
GSFC University
Vadodara
saurabh_er@rediffmail.com

Article Info
Page Number: 1483-1495
Publication Issue:
Vol.71 No.3 (2022)

Article History
Article Received: *12 January 2022*
Revised: *25 Febuary 2022*
Accepted*: 20 April 2022*
Publication*: 09June 2022*

Abstract

Named Entity Recognition (NER) is a method to search for a particular Named Entity (NE) from a file or an image, recognize it and classify it into specified Entity Classes like Name, Location, Organization, Numbers and Others Categories. It is the most useful element of the technique knownas Natural Language Processing (NLP) which makes text extraction very easy. This paper is about Named Entity Recognition (NER) for Gujarati language. Not much work has been done in NER for Gujarati. There is no standard dataset available for Gujrati NER. Hence we have created two datasetsfor NER in Gujrati. In this paper, an NER tagger is build usingSpacy. The NER tagger is trained with 100% accuracy and capable of identifying person, location and organization names.From the news headlines dataset the NER tagger is able to identify the named entities for entertainment, business and technical etc.

Keywords: NER, Gujrati NER, Spacy

I. INTRODUCTION

One of the key components of most successful NLP applications is the Named Entity Recognition (NER) modulewhich accurately identifies the entities in text such as date, time, location, quantities, names and product specifications. There are already existing sophisticated systems for NER such as spaCy, Stanford NER, etc. but most of them are built withgeneral purpose for a wide range of NLP applications such as Information Retrieval, Document classification and otherapplications of unstructured data analysis. Named entity recognition (NER)is probably the first step towardsinformation extraction that seeks to locate and classify namedentities in text into pre-defined categories such as the names of persons, organizations,

locations, expressions of times, quantities, monetary values, percentages, etc. NER is used in many fields in Natural Language Processing (NLP), and it can help answering many real-world questions, such as:

- Which companies were mentioned in the newsarticle?
- Were specified products mentioned in complaints or reviews?

Does the tweet contain the name of a person? Does the tweet contain this person's location? Named Entities are one of the most important textual unit in the Information Extraction domain as they express an important part of the meaning of a document. NER is also animportant part of Natural Language Processing (NLP) applications like Machine Translation and Text Summarization and is also used in search engines. For Gujarati language no such NER tagger exist and hence this paper propose a NER tagger for Gujrati text to solve the problem. A remarkable amount of work has been carried outfor many languages like English, Greek, Chinese etc. But, stilla wide scope is open for Indian Origin Languages like Hindi,Gujarati, Devanagari etc. As Gujarati is not only the Indian Language, but a language that is most spoken in Gujarat. Thus, in this paper, we emphasis on proposing a NER based schemefor Gujarati Language.

II. RELATED WORK

Aiming at the problem of Khmer named entity recognition, authors in [1] proposed a method fusing Khmer entity characteristics based on the universal feature templates. For the relatively stable entity that is formed of time expressionsand digital expressions, authors recognize it using artificial rules; For the complex entity that is formed of names, locations, and organizations, authors use Conditional RandomFields algorithm, taking word, part of speech, contextual information and Khmer entity characteristics into consideration, to build a complex entity recognition model to recognize it. Experimental results show that the named entity recognition method fusing Khmer entity characteristics has abetter effect.

Authors in [2] have presented the evaluation of Named EntityRecognition task for a resource poor language like Punjabi. Different challenges posed in the task have also been discussed. A annotated corpus of 2,00,000 words have been developed for recognition task. The system has been evaluatedon different machine learning models like Hidden Markov Model, Maximum Entropy and Conditional Random Fields with f-score values of 77.61, 83.65 and 93.21 respectively. The purpose of this research in [3] is to reduce humans as annotation effort for clinical notes, to improve consistency, and to decrease cost of annotation. The aim of this research [3] is to annotate clinical texts to extract biomedical namesand terms. Authors in [3] have done unsupervised and semi-supervised Named Entity Recognition (NER) through exact matching in UMLS. The data sets that have been used wereprovided by SemEval 2015 (task 14) natural languageprocessing competition, including 199 clinical notes intraining set and 133 notes in test set. With this method authorsgot 60% of f-score, and training files for next process (trainingCRFs) were generated. The second step involves usingConditional Random Fields (CRFs). The results generated inthe first step were used to train the CRF. CRFs learn fromtraining data the general contexts in which named entitiesoccur. By supervised learning authors got 73% F-score whilewe got 62.7% by the proposed unsupervised approach. Authors in [4] critically compares the state-of-the-arts, meritsand gaps in the works. Review reveals the

need to focus ontechniques to increase accuracy of symbol spotting.

Aiming at the problem of fuzzy entity recognition and less labeled data in the field of traditional Chinese medicine, a named entity recognition model based on Bert-BiLSTM-CRF is constructed and tested on the corresponding data set in [5].According to the text information, it is divided into five typesof entities: symptoms, disease names, time, prescription names, and drug names. The results show that the model hasthe highest accuracy in identifying drug names. In order to further prove the superiority of this model, three groups of control groups composed of other models are set up.

In [6], it is expressed how to integrate ontologies to NER problem as a solution proposal. Authors implemented the overall operation by initially gathering essential Turkish entities, fetching the linked information from the ontologies about entity names, extracting features with this linked information and classifying the entities according to these features.

Authors in [7] first briefly introduces the developing processof named entity recognition, and describes the basic concept, objectives and difficulties of named entity recognition. Then, it summarizes the methods of named entity recognition from based on rules and dictionary method, based on statistical method, to the deep learning method and migration of learning, and looks forward to the future development of named entity recognition.

DATASET PREPARATION FOR GUJRATHI NAMED ENTITY RECOGNITION

During this work of Gujrati named entity recognition, we have used two datasets.

A. Custom Dataset of Gujratho words

Since no standard dataset is available for Gujrati named entity recognition, we have prepared the dataset of Gujrathi Keywords as depicted in table 1

TABLE I. CUSTOM DATASET OF GUJRATHI KEYWORDS

Category	***Number of Gujrati Keywords***	***Sample Wods***
Cities	34	ગાાંધીધામ આણાંદ નવસારી મોરબી નડિયાદ સુરન્દ્ે રનગર ભરૂચ મહેસાણા.
Countries	195	સીડરયા તાડિડિસ્તાન તાન્દ્ઝાડનયા થાઈલન્દ્િ ડતમોર-લેસ્તે જાઓ ટોાંગા
		ગુરુવાર શુક્રવાર શડનવાર જાન્યુઆરી ફેબ્રુઆરી િુચ એડિલ શિ છે

Days-Months	19	
Names	100	Rી નારાયણ તટુ રાણે Rી નરન્દ્ે ર મોદી Rી સબાાનદાં સોનોવાલ Rી મુખ્તાર અબ્બાસ નિવી િો.વીરન્દ્ે ર િુમાર Rી ડગડરરાિ ડસાંહ Rી જ્યોડતરાડદત્ય એમ. ડસાંડધયા Rી અડિની વૈષ્ણવ Rી રામચાંર િસાદ ડસાંહ Rી પશુપડત િુમાર પારસ
Numbers	37	૧ એિ ek ૨ બે be ૩ ત્રણ traṇ ૪ ચાર chaar ૫ પાાંચ pããch ૬ છ chha ૭ સાત saat ૮ આઠ aaṭh ૯ નવ nav ૧૦ દસ das
		એલેડબબિ આલ્ફા લાવલ ઓલસેિ ટેક્નોલોજીસ આલોિ ઇન્દ્ર્સ્ટર ીઝ અલ્સ્ટોમ િોિક્ટે ્સ ઇડન્દ્ર્યા અનાંત રાિ ઇન્દ્ર્સ્ટર ીઝ અમરા રાજા બેટરીઝ એબટેિ ઈડન્દ્ર્યા એએનજી ઓટો અાંસલ હાઉડસાંગ એન્દ્ર્િ ન્દ્ર્સ્ટર ક્ટશન

Organization	320	
State	28	િેરળ મધ્યિદેશ મહારાષ્ટ્ર મડણપુર મેઘાલય

Category	***Number of Gujrati Keywords***	***Sample Wods***
		ડમઝોરમ નાગાલેન્દ્ ઓડિશા પજા બ રા સ્થાન
Total	733	

We have prepared Custom dataset of Gujrati keywords with 7 categories and 733 Gujrati words.We have created a JSON file of all keywords with tagging.

B. Gujrathi News Classification Dataset

For preparing this dataset we have used Gujrati Kaggle NewsClassification Dataset [8]. The dataset contains 6500 news articles headlines. News headlines are divided into threecategories as depicted in table 2.

TABLE II. KAGGLE GUJRTHI NEWS CLASSIFICATION DTASET

Category	***Number of Gujrati news***	***Sample News***
Entertainment	2134	શાહરૂખની સાથે ફિલ્મમા કામ કરી ચુકે છે અનુ કાથી ઠપકો ખાધેલો વ્યક્તિ સલમાન સાથે પગ ા બાદ નથી મળિો 'સોરી' અને 'થેન્ક્યૂ' નો ચાન્સ

Business	2358	ચનનો શાધ ાઈમા આઈસી આઈસીઆઇ બેંક ખોલશે ેની નવી બ્ાચ સકુ ્યા યોજનામા 3 મોટા બદલાવ, 1000 નહીં હવે 250 ૂપયામા ખલુ શે ખાતુ
Technical	2008	આજે અહીથી ખરીદો OnePlus 6, મળશે 25,000 ૂપયા સધુ ીનો ાયદો Ford Aspire CNG ભાર મા લોન્ચ, જાણી લો ફકંમ
Total	6500	

We have prepared the dataset of Gujrati named entity recognition by annotating the keywords from news classification dataset. Some of the annotations are depicted intable 3.

TABLE III. CUSTOM ANNOTATIONS OF GUJRTHI NEWS CLASSIFICATION DTASET

Category	***Sample annotated News***
	["'સેલ્ફી શોખીનો માટે લોન્ચ થયો Huawei Y6 Pro, જાણો ફિચસસ' ,",{"entities":[[1 ,8,"ENTERTAINMENT"],[9,15,"ENTERTAIN MENT"],[21,26,"BUS INESS"],[31,44,"TE CHNICAL"],[51,57," TECHNICAL"]]}]
	["'જજયોની ચગફટ- હવે યઝુ સસ JioPhone દ્વારા બકુ કરી શકશે ટ્રેનની ફટફકટ' ,",{"entities":[[1 ,8,"TECHNICAL"],[9 ,15,"ENTERTAINMENT "],[20,26,"TECHNIC

	AL"],[27,35,"TECHN
Entertainment	ICAL"]]}]
	["'10 હજાર રૂપપયાથી પણ ઓછામાાં મળશે બ્ાન્ડેડ smart LED TV, આજે અંપિમ િક'
	,",{"entities":[[1
	0,18,"BUSINESS"],[
	29,33,"BUSINESS"],
	[34,43,"BUSINESS"]
	,[44,49,"TECHNICAL "],[50,56,"BUSINES
	S"]]}]
	["'Redmi Note 7માાં હોઈ શકે છે આ ખાસ
	ીચસસ, જાણો ફકિમ'
	,",{"entities":[[1
	,7,"TECHNICAL"],[8
	,17,"TECHNICAL"],[35,41,"TECHNICAL"]

Business	,[48,53,"BUSINESS"]]}]
Technical	['"દુ પનયાભરમાાં િેસબ્કુ ડાઉન, હકે સસ અટેક ન હોવાનો કાંપનીનો દાવો' ,",{"entities":[[1 3,19,"TECHNICAL"], [26,32,"TECHNICAL"]

Category	***Sample annotated News***
	['"ભાર િમા લોચ થયા Galaxy S પસરીઝના આાત્રણ મોટ િોન, મળશે જબરદ િ કેશબેક' ,",{"entities":[[2 0,28,"TECHNICAL"], [44,54,"TECHNICAL"]]}]

IV. NAMED ENTITY RECOGNITION USING SPACY

SpaCy is an open-source library for advanced Natural Language Processing in Python. It is designed specifically for production use and helps build applications that process and "understand" large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning. Some of thefeatures provided by spaCy are- Tokenization, Parts-of- Speech (PoS) Tagging, Text Classification and Named EntityRecognition.

SpaCy provides an exceptionally efficient statistical system for NER in python, which can assign labels to groupsof tokens which are contiguous. It provides a default model which can recognize a wide range of named or numerical entities, which include *person, organization, language, eventetc.* Apart from these default entities, spaCy also gives us theliberty to add

arbitrary classes to the NER model, by training the model to update it with newer trained examples.

A. Installation

SpaCy can be installed using a simple pip install. You will also need to download the language model for the language you wish to use spaCy for.

pip install -U spacy python -m spacy download en

B. Data Preprocessing

SpaCy requires the training data to be in the following format-

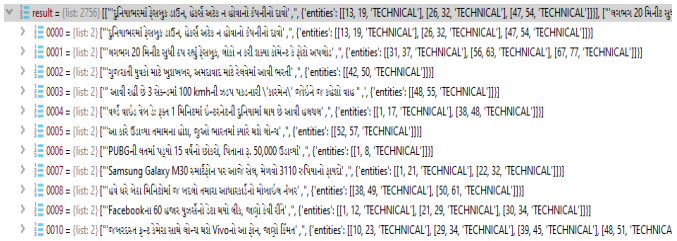

Figure 1: Data preprocessing for Spacy

So we have to convert our data which is in .csv format to the above format. *(There are also other forms of training data which spaCy accepts. Refer the documentation for more details.)* We first drop the columns Sentence # and POS as we don't need them and then convert the .csv file to .tsv file.

Next, we have to run the script below to get the training data in .json format.

C. *first we import spacy and proceed.*

```
import spacy
nlp = spacy.blank("gu") # load a new spacymodel
```

D. then we imort dataset file and load data in TRAIN_DATA

```
import jsonf =
open('news_classification.json',encoding="utf=8")
TRAIN_DATA = json.load(f)
```

E. DocBin object is created and all data is stored in spacy model. You have to add labels from training data to the ner using ner.add_label() method of pipeline. Below code demonstrates the same

```
# Adding labels to the `ner`
for _, annotations in TRAIN_DATA:
for ent in annotations.get("entities"):ner.add_label(ent[2])
```

F. To train ner model, the model has to be looped over the example for sufficient number of iterations. If you train it for like just 5 or 6 iterations, it may not be effective.

Before every iteration it's a good practice to shuffle the examples randomly through random.shuffle() function .

G. Training of our NER is complete now. You can test if the ner is now working as you expected. If it's not up to your expectations, include more training examples and try again.

doc2 = nlp_ner("દીપ◌કાએ વોગ મગઝીન મ◌ાટે ક◌રાવ્◌ ફ◌ોટ◌ોશૂટ , જ◌ઓ હોટ Pics એક ઇવસનની ર◌ાહત બ◌ાદ ફ◌રી ◌ૂટે ◌ોલના ભ◌ાવમ◌ાું ભડક◌ો, ર◌ૂ. 89.60 પ્રતિ લિટર")
spacy.displacy.render(doc2, style="**ent**", jupyter=**True**) *# display in Jupyter*

ENTERTAINMENT દીપકાએ
BUSINESS ભાવમા

We use python's spaCy module for training the NER model.spaCy's models are **statistical** and every "decision" they make — for example, which part-of-speech tag to assign, or whether a word is a named entity — is a **prediction**. This prediction is based on the examples the model has seenduring **training**.

The model is then shown the unlabelled text and will make aprediction. Because we know the correct answer, we can givethe model feedback on its prediction in the form of an **error gradient** of the **loss function** that calculates the difference between the training example and the expected output. The greater the difference, the more significant the gradient and the updates to our model.

When training a model, we don't just want it to memorise ourexamples — we want it to come up with theory that canbe **generalised across other examples**. After all, we don't just want the model to learn that this one instance of "Amazon" right here is a company — we want it to learn that"Amazon", in contexts *like this*, is most likely a company. In order to tune the accuracy, we process our training examplesin batches, and experiment with minibatch sizes and dropoutrates.

V. EXPERIMENTS

A. *Custom Dataset of Gujrthi Words*

We have created the spacy object for Gujrthi NER and trained the model using custom pipeline. We have evaluated the performance of the model using precision, recall, f measure and accuracy. We achieved 100% performance for training.We have evaluated the performance of the model and resultsare as shown in figure 2.

Figure 2: Named Entity Recognition of Custom Gujrati Words Dataset

B. Gujrthi News Dataset

Gujrthi News Classification dataset from from Kaggle [8] istagged for entities. A Spacy model is trained for the dataset and the results are depicted in figure 3.

We have evaluated the performance of the model usingprecision, recall, f measure and accuracy. We achieved 100% performance for training.

Figure 4: Named Entity Recognition for tagged Gujrthi NewsDataset

VI. CONCLUSION

Named entity recognition is one of the important topics in theresearch area of natural language processing. Named entity recognition studies conducted on Gujrati texts are quite limited, compared to the studies on other languages. Besides,the lack of common data sets makes the comparison ofdifferent approaches harder. In this study, we have constructed two datasets for Gujrati NER. Custom NERdataset of Gujrati keywords is prepared consisting of keywords from places, names, states, countries, organization and days-months. Another dataset comprising news articles in Gujrati annotated with named entities is presented. The annotations comprise the basic named entity types of business, entertainment and technical names. Additionally, tobe used as reference in future studies, a named entity recognition system using spacy is evaluated on the final formof this data set and the corresponding evaluation results are presented. It is envisioned that our study will contribute to the advancement of named entity recognition studies on Gujrati texts.

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word “acknowledgment” inAmerica is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R.

C. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] H. Pan, X. Yan, Z. Yu and J. Guo, "A Khmer named entity recognition method by fusing language characteristics," *The 26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, China, 2014, pp. 4003-4007.

[2] A. B. YILMAZ, Y. S. TASPINAR, and M. Koklu, “Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms”, Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 269–274, May 2022.

[3] A. Kaur and S. Khattar, "A systematic exposition of Punjabi Named Entity Recognition using different Machine Learning models," *2021 Third International Conference on Inventive*

Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1625-1628.

[4] O. Ghiasvand and R. J. Kate, "Biomedical Named Entity Recognition with less Supervision," *2015 International Conference on Healthcare Informatics*, Dallas, TX, USA, 2015, pp. 495-495.

[5] Gill, D. R. . (2022). A Study of Framework of Behavioural Driven Development: Methodologies, Advantages, and Challenges. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 09–12. https://doi.org/10.17762/ijfrcsce.v8i2.2068

[6] O. KaiBin and Y. K. Hooi, "Critical Literature Review of Named Entity Recognition in Symbol Spotting," *2021 International Conference on Computer & Information Sciences (ICCOINS)*, Kuching, Malaysia, 2021, pp. 220-225.

[7] Hoda, S. A. ., and D. A. C. . Mondal. “A Study of Data Security on E-Governance Using Steganographic Optimization Algorithms”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 5, May 2022, pp. 13-21, doi:10.17762/ijritcc.v10i5.5548.

[8] Q. Qu, H. Kan, Y. Wu and Y. Gao, "Named Entity Recognition of TCMText Based on Bert Model," *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Hefei, China, 2020, pp. 652-655.

[9] Deepak Mathur, N. K. V. . (2022). Analysis & Prediction of Road Accident Data for NH-19/44. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 13–33. https://doi.org/10.17762/ijrmee.v9i2.366

[10] O. Çiftçi and F. Soygazi, "Ontology Based Turkish Named Entity Recognition," *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey, 2020, pp. 1-5.

[11] Q. Guo, S. Wang and F. Wan, "Research on Named Entity Recognition for Information Extraction," *2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*, Manchester, United Kingdom, 2020, pp. 121-124.

[12] Kaggle Gujrati News Dataset: https://www.kaggle.com/datasets/disisbig/gujarati-news-dataset