Named Entity Recognition from Gujarati Text Using Rule-Based Approach

Dikshan N. Shah[1(✉)] and Harshad B. Bhadka[2]

[1] Faculty of Computer Applications, S S Agrawal Institute of Computer Science, Navsari, India
dikshan817@gmail.com
[2] Faculty of Computer Science, C U Shah University, Wadhwan, India
harshad.bhadka@yahoo.com

Abstract. NER which is known as Named Entity Recognition is an application of Natural Language Processing (NLP). NER is an activity of Information Extraction. NER is a task used for automated text processing for various industries, a key concept for academics, artificial intelligence, robotics, Bioinformatics and much more. NER is always an essential activity when dealing with chief NLP activity such as machine translation, question-answering, document summarization etc. Most NER work has been done for other European languages. NER work has been done in few Indian constitutional languages. Not enough work is possible due to some challenges such as lack of resources, ambiguity in language, morphologically rich and much more. In this paper, to identify various named entities from a text document, rules are defined using Rule-based approach. Based on defined rules, three different test cases computed on the training dataset and achieved 70% of accuracy.

Keywords: NER · Rule-based approach · Constitutional languages
Tagset · Tithi

1 Introduction

The phrase Named Entity (NE) was coined during the 6th Message Understanding Conference (MUC-6) in 1995. Many NER systems were developed after that. Foremost work has been done in European languages and all systems were highly precise [6]. Named Entity is the structured information mentioning to predefined proper names like persons, locations, and organizations, year, date, month, monetary amounts, percentages as well as temporal and numeric expressions from text [2].

Named Entity Recognition (NER) systems proved to be very significant for many tasks in Natural Language Processing (NLP) such as information retrieval, machine translation, information extraction, question answering systems. The objectives of NER is to classify each word of a document into predefined target named entities classes.

A. Abraham et al. (Eds.): ISDA 2017, AISC 736, pp. 797–805, 2018.
https://doi.org/10.1007/978-3-319-76348-4_76

1.1 Existing NER Approaches

Present NER systems have been built using mainly knowledge-based or linguistic, and machine learning approach.

1.1.1 Rule-Based Approach

The linguistic approach or Knowledge-based approach is basically called as a rule-based approach which uses a set of hand-crafted rules deliberate and described by human experts, especially linguists. This approach considers a set of patterns containing grammatical, syntactic, linguistic and orthographic features in a grouping with dictionaries. It is a prerequisite to have a thorough knowledge of target language as it is a time-consuming task to develop such kind of system.

1.1.2 Statistical Approach

Machine Learning or a Statistical approach is a swift way to build an NER system which fundamentally supports rule-based systems or use sequence labeling algorithms to collect knowledge from a collection of training examples. The accuracy of this approach is purely dependent upon the training dataset. Various Machine Learning models used for NER systems like Hidden Markov Model, Conditional Random Field, and Maximum Entropy.

1.1.3 Hybrid System

Use of Statistical tools as well as linguistic rules and combinations of both approaches make a system more precise and effective.

1.2 About the Gujarati Language

Basically, Among the Indo-European language family, Gujarati is well-known Indo-Aryan language and it was tailored from the Devanagari script. Alphabets of this language mainly include 34 consonants and 14 vowels. [1] A language is very widespread and spoken by more than 50 million people across the India. It is the official language of the Gujarat state of India.

2 Related Work on Different Indian Languages in NER

Among the constitutional Indian languages, NER work has been done in some languages. NER approaches used in various Indian languages with their accuracies are mentioned in Table 1 as follows:

Table 1. Different approaches used for various Indian Languages according to their accuracies

Author	Language	Method/Approach	Precision	Recall	F-Measure	Accuracy
[1]	Gujarati	Inflectional stemmer	–	–	–	90.7%
[1]	Gujarati	Derivational stemmer	–	–	–	70.7%
[2]	Hindi	Rule based	75.86%	79.17%	77.48%	–
[3]	Kannada	Rule based	78.6%	77.22%	77.2%	–
[4]	Malay	Rule based	85%	94.44%	89.47%	–
[5]	Kannada	Hybrid	–	–	–	94.85%
[8]	Hindi	Rule based	–	–	–	79.06%
[9]	Dogri	Rule based	–	–	–	90%
[10]	Hindi	HMM	–	–	–	98.37%
[11]	Tamil	CRF	–	–	–	87.20%
[11]	Tamil	SVM	–	–	–	86.06%
[12]	Hindi	Hybrid	–	–	–	95.77%
[13]	Arabic	Rule based	92.25%	91.25%	91.71%	–
[14]	English	Rule based	–	–	–	88.19%

3 Rule-Based Approach

A morphological analyzer for the Hindi language analyze Hindi sentences and produce its features with its root words. [7] As Rule-based approach is a domain specific, rules define for one language will not apply for other languages. Some Rules used to identify different tags in the Gujarati language are as follows:

3.1 Date and Time

This Rule is applied on given input which contains the various date and Time formats. Regular Expressions are used to identify these kinds of tags [15]. Following are date and time tagset examples:

3.1.1 Year વિક્રમસંવત, ઇસવિસન, વર્ષ, સાલ

- 'સંવત' *(Samvat)* refers to the epoch of the several Hindu calendar systems in India and also in Nepal. There are three most significant 'સંવત' *(Samvat)*: Vikrama era, Old Shaka era and Shaka era of 78 AD [15].

3.1.2 Month Names

- જાન્યુઆરી ફેબ્રુઆરી માર્ચ એપ્રિલ મે જૂન જુલાઈ ઓગસ્ટ સપ્ટેમ્બર ઓક્ટોબર, નવેમ્બર, ડિસેમ્બર

- The names of the Indian months diverge by region. Hindu calendars are based on lunar cycle and usually phonetic variants of each other.
- કારતક, માગશર, પોષ, મહા, ફાગણ, ચૈતર, વૈશાખ, જેઠ, અષાડ, શ્રાવણ, ભાદરવો, આસો [15].

3.1.3 Days

- સોમવાર, મંગળવાર, બુધવાર, ગુરુવાર, શુક્રવાર, શનિવાર, રવિવાર

The Hindu calendar has two measures of a day, one based on the lunar movement and the other on solar. The solar day or civil day is called divas (દિવસ), and the lunar day is called tithhi (તિથિ). A lunar month has 30 tithhi. Lunar month starts with Kartak (કારતક).
પ્રતિપદા, દ્વિતીયા, તૃતીયા, ચતુર્થી, પંચમી, શ્રૃષ્ટિ, સપ્તમા, અષ્ટમી, નવમી, દસમી, એકાદસી, દ્વાદસી, ત્રાયોદશી, ચતુર્દસી, પૂર્ણિમા, અમાવસ્યા, એકમ, બીજ, ત્રિજ, ચોથ, પંચમ, છઠ્ઠ, સાતમ, આઠમ, નોમ, દસમ, અગિયારસ, બારસ, તેરસ, ચૌદસ, અમાસ, પૂનમ [16].

3.2 Location

Suffix matching is used for types of location names and terms. Different suffix makes different location names of Indian States and Cities are as follows: [17]

- Location names that end with ‘pure’ (પુર) i.e. - રામપુર, સુંદરપુર, જયપુર, ઉદયપુર
- Location names that end with ‘Ghar’ (ગઢ) i.e. - રાયગઢ, ચંદીગઢ, સોનગઢ
- Location names that end with ‘stan’ (સ્તાન) i.e. - હિન્દુસ્તાન, પાકિસ્તાન, અફઘાનિસ્તાન
- Location names that end with ‘bad’ (બાદ) i.e. – અમદાવાદ, ફરિદાબાદ, હૈદરાબાદ
- Location names that end with ‘Nagar’ (નગર) i.e. – શ્રીનગર, ગાંધીનગર, ગંગાનગર
- Location names that end with ‘pat’ (પત) i.e. - પાણીપત, સોનિપત
- Location names that end with ‘nath’ (નાથ) i.e. - કેદારનાથ, સોમનાથ, બદ્રીનાથ
- Location names that end with ‘mer’ (મેર) i.e. - જેસલમેર, બારમેલ
- Location names that end with ‘kot’ (કોટ) i.e. - રાજકોટ, પઠાણકોટ, સિયાલકોટ
- Location names that end with ‘Ishwar’ (ઇશ્વર) i.e. - અંકલેશ્વર, મહાબળેશ્વર, ભુવનેશ્વર
- Location names that end with ‘Wada’ (વાડા) i.e. - બામનવાડા, ભિલવારા, તેલવાડા
- Location names that end with ‘giri’ (ગિરિ) i.e. – રત્નાગિરિ, ચંદ્રગીરી, ધૌલગીરી
- Location names that end with ‘Puram’ (પુરમ) i.e. – તિરુવનંતપુરમ, મલ્લાઇપુરમ
- Location names that end with ‘uru’ (ઉરુ) i.e. – બેંગલુરુ, મેંગલુરુ
- Location names that end with ‘patnam’ (પટનમ) i.e. – વિશાખાપટ્ટનમ, માસુલિપટનમ
- Location names that end with ‘guri’ (ગુડી) i.e. – સિલિગુડી, જલપાઇગુડી, મેનાગુરી
- Location names that end with ‘tal’ (તાલ) i.e. – નૈનિતાલ
- Location names that end with ‘Dwar’ (દ્વાર) i.e. – હરિદ્વાર, કોટવાર
- Location names that end with ‘Wada’ (વાડા) i.e. – ભીલવાડા, બામનવાડા

- Location names that end with 'Palli' (પલ્લી) i.e. – તિરુચિરાપલ્લી, જલાહલ્લી
- Location names that end with 'Malai' (માલાઈ) i.e. – કોલામાલાઈ, અન્નામલાઈ,, સબરીમાલાઈ

3.3 To Identify Some Abbreviations in Date and Time Tag Entities

Abbreviations point to an original name. Some words used in their abbreviated form for a date, month and year entities. Examples: ઇ.સ., વિ.સ., તા., જાન્યુ, ફેબ્રુ., એપ્રિ., જૂ., ઓગ., સપ્ટે.,ઓક્ટો.,નવે.,ડિસે., સોમ, મંગળ, બુધ, ગુરુ, શુક, શનિ, રવિ [15].

3.4 For Numerals

There is a difference between mentioning numbers. Two types of number system we used: Hindu Arabic Numerals and Gujarati Numerals. Numbers have different number names in different languages. Number 0 to 100 written in both format is different and their Gujarati names also [18] (Table 2).

Table 2. Number Names of Hindu Arabic Numerals in Gujarati

Hindu-Arabic Numeral	Gujarati numeral	Gujarati name	Hindu-Arabic numeral	Gujarati numeral	Gujarati name
0	૦	શૂન્ય	31	૩૧	એકત્રીસ
1	૧	એક	32	૩૨	બત્રીસ
2	૨	બે	33	૩૩	તેત્રીસ
3	૩	ત્રણ	34	૩૪	ચોત્રીસ
4	૪	ચાર	35	૩૫	પાંત્રીસ
5	૫	પાંચ	36	૩૬	છત્રીસ
6	૬	છ	37	૩૭	સડત્રીસ
7	૭	સાત	38	૩૮	અડત્રીસ
8	૮	આઠ	39	૩૯	ઓગણચાલીસ
9	૯	નવ	40	૪૦	ચાલીસ
10	૧૦	દસ	41	૪૧	એકતાલીસ
11	૧૧	અગિયાર	42	૪૨	બેતાલીસ
12	૧૨	બાર	43	૪૩	ત્રેતાલીસ
13	૧૩	તેર	44	૪૪	ચુંમાલીસ
14	૧૪	ચૌદ	45	૪૫	પિસ્તાલીસ
15	૧૫	પંદર	46	૪૬	છેતાલીસ
16	૧૬	સોળ	47	૪૭	સુડતાલીસ
17	૧૭	સત્તર	48	૪૮	અડતાલીસ
18	૧૮	અઢાર	49	૪૯	ઓગણપચાસ

19	૧૯	ઓગણિસ	50	૫૦	પચાસ
20	૨૦	વીસ	51	૫૧	એકાવન
21	૨૧	એકવીસ	52	૫૨	બાવન
22	૨૨	બાવીસ	53	૫૩	ત્રેપન
23	૨૩	તેવીસ	54	૫૪	ચોપન
24	૨૪	ચોવીસ	55	૫૫	પંચાવન
25	૨૫	પચ્ચીસ	56	૫૬	છપ્પન
26	૨૬	છવીસ	57	૫૭	સત્તાવન
27	૨૭	સત્તાવીસ	58	૫૮	અઠ્ઠાવન
28	૨૮	અઠ્ઠાવીસ	59	૫૯	ઓગણસાઠ
29	૨૯	ઓગણત્રીસ	60	૬૦	સાઈઠ
30	૩૦	ત્રીસ	81	૮૧	એક્યાસી
61	૬૧	એકસઠ	82	૮૨	બ્યાસી
62	૬૨	બાસઠ	83	૮૩	ત્યાસી
63	૬૩	ત્રેસઠ	84	૮૪	ચોર્યાસી
64	૬૪	ચોસઠ	85	૮૫	પંચાસી
65	૬૫	પાંસઠ	86	૮૬	છ્યાસી
66	૬૬	છાસઠ	87	૮૭	સિત્યાસી
67	૬૭	સડસઠ	88	૮૮	ઈઠ્યાસી
68	૬૮	અડસઠ	89	૮૯	નેવ્યાસી
69	૬૯	અગણોસિત્તેર	90	૯૦	નેવું
70	૭૦	સિત્તેર	91	૯૧	એકાણું
71	૭૧	એકોતેર	92	૯૨	બાણું
72	૭૨	બોતેર	93	૯૩	ત્રાણું
73	૭૩	તોતેર	94	૯૪	ચોરાણું
74	૭૪	ચુમોતેર	95	૯૫	પંચાણું
75	૭૫	પંચોતેર	96	૯૬	છન્નું
76	૭૬	છોતેર	97	૯૭	સત્તાણું
77	૭૭	સિત્યોતેર	98	૯૮	અઠ્ઠાણું
78	૭૮	ઇઠ્યોતેર	99	૯૯	નવ્વાણું
79	૭૯	ઓગણાએંસી	100	૧૦૦	સો
80	૮૦	એંસી	1000	૧૦૦૦	હજાર

4 Research Methodology

We have developed various rules using a rule-based approach which helps to recognize various named entities. For Identification of Named Entity, we have collected document in the Gujarati language as a corpus from E-newspaper '**Gujarat Samachar**'. There are various categories of news as Entertainment, Sports, Religious and much more. Among them, we have gathered 100 sports category documents to identify various Named Entity tagset.

A. Preparation of Database

Based on various categories of tagset, following dictionaries are created.

Date Dictionary: Date tagset contains Day, Month number and name, and Year. The day is also categorized based on Hindu calendar and Panchang (પંચાંગ). Tithis (તિથિ) and days (દિવસો) stored in gazetteer list.

Location Dictionary: Here Location names are only within a limited range of area or for a specific country. 21 Suffix stripping rules are created for Location Names as City or State or Village names of India.

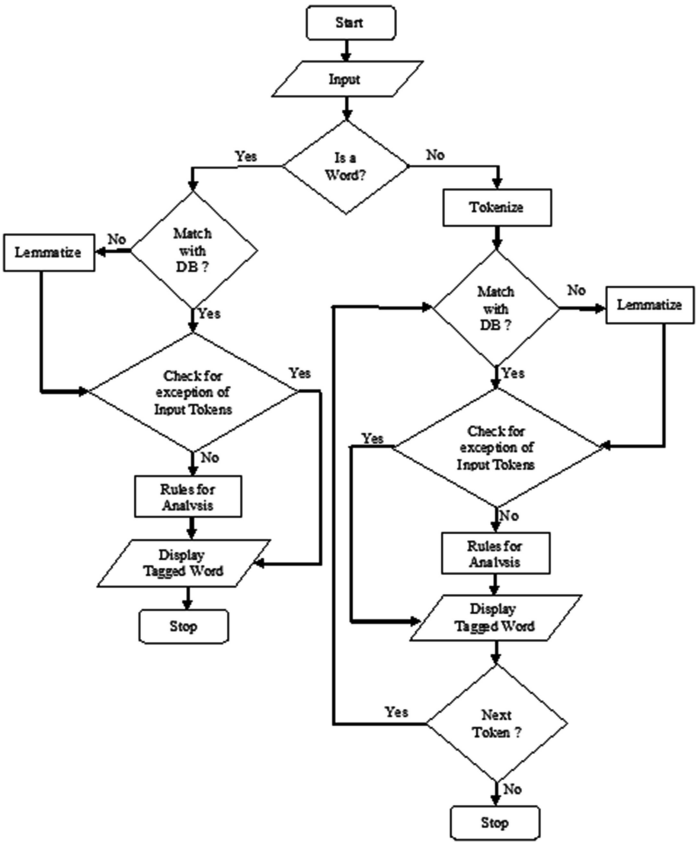

Fig. 1. Flowchart for Rule based NER for Gujarati language

Abbreviation Dictionary: Various abbreviations of date, month, day are listed in it.
Number-names dictionary: Based on Hindu-Arabic numerals, Gujarati number names listed for 0 to 100 digits (Fig. 1).

B. **Architecture of System**

Step – 1 Input text - Through file upload, upload a file which is a text file comprising raw data in the Gujarati language.
Step – 2 Preprocessing – Prepare Gujarati text document for preprocessing.
Step – 3 Tokenization - Input text tokenized word by word for pattern matching.
Step – 4 Entity detection - Detection of Date, Time and Location entities based on created rules and if any rules matched go to step 7.
Step – 5 Detection of **Abbreviation Names**, if matches are found and compared with gazetteer list, go to step 7.
Step – 6 Detection of **Numbers and its Number names** from Non-numerals practice and if found any matches then go to step 7.
Step – 7 Display **tagged output** generated by the system with the untagged result to the user.
Step – 8 End

5 Experimental Result and Analysis

The core objectives of such experiment are to identify the kinds of patterns of named entities by the proposed NER algorithm. We have collected documents of Sports category to recognize various Named Entities such as date, Day names, Month Names, Tithhi (તિથિ), Location and numerals (Table 3).

Table 3. Apply various test cases on dataset

Accuracy	Test No	Tagset	No. of Words	Correctly observed tag
	Test 1	Date, Days, Tithi	147	97
	Test 2	Months	129	68
	Test 3	Locations	87	29

Among the given 363 words, 194 entities are correctly identified by applying various test cases and achieved 70% of accuracy.

6 Conclusions

An innovative technique can be build up to develop the performance of NER in the Gujarati language. We have developed rules to identify various named entities which is a very beneficial in many significant applications. We have studied various existing approaches of NER and analyzed that among the various constitutional Indian

languages, lots of scopes is for NER in Indian languages. By implementing various rules on given dataset we attained 70% of accuracy. As a future work, we can build more precise rules for much more named entities to achieve good accuracy.

References

1. Athavale, V., Bharadwaj, S., Pamecha, M., Prabhu, A., Shrivastava, M.: Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Scarcity (2016)
2. Jiandani, K.S.D., Bhattacharyya, P.: Hybrid inflectional stemmer and rule-based derivational stemmer for Gujarati. In: Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011), November 2011
3. Amarappa, S., Sathyanarayana, S.V.: Kannada named entity recognition and classification (nerc) based on multinomial naïve Bayes (MNB) classifier. Int. J. Nat. Lang. Comput. (IJNLC) **4**, 39–52 (2015)
4. Alfred, R., Leong, L.C., On, C.K., Anthony, P.: Malay named entity recognition based on rule-based approach. Int. J. Mach. Learn. Comput. **4**(3), 300–306 (2014)
5. Sathyanarayana, S.A.: A hybrid approach for named entity recognition, classification and extraction (NERCE) in Kannada documents. In: Proceedings of International Conference on Multimedia Processing, Communication, and Info. Tech., MPCIT (2013)
6. Singh, A.K.: Named entity recognition for south and south east asian languages: taking stock. In: Proceedings of the IJCNLP Workshop on NER for South and South East Asian Languages, pp 5–16 (2008)
7. Agarwal, A., Singh, S.P., Kumar, A., Darbari, H.: Morphological analyser for hindi-a rule-based implementation. Int. J. Adv. Comput. Res. **4**(1), 19 (2014)
8. Sharma, L.K., Mittal, N.: Named entity based answer extraction from hindi text corpus using n-grams. In: 11th International Conference on Natural Language Processing, p. 362, December 2014
9. Sasan, T.S., Jamwal, S.S.: Transliteration of name entities using rule-based approach. Int. J. Adv. Res. Comput. Sci. Soft. Eng., **6**(6) (2016)
10. Jahan, N., Morwal, S., Chopra, D.: Named entity recognition in Indian languages using gazetteer method and hidden Markov model: a hybrid approach. IJCSET, March 2012
11. Abinaya, N., Kumar, M.A., Soman, K.P.: Randomized kernel approach for named entity recognition in Tamil. Indian J. Sci. Technol. **8**(24), 1–7 (2015)
12. Kaur, Y., Kaur, E.: Named Entity Recognition system for Hindi Language using a combination of rule-based approach and list lookup approach. Int. J. Sci. Res. Manag. (IJSRM) **3**(3), 2300–2306 (2015)
13. Aboaoga, M., Ab Aziz, M.J.: Arabic person names recognition by using a rule-based approach. J. Comput. Sci. **9**(7), 922 (2013)
14. Bhalla, D., Joshi, N., Mathur, I.: Rule-based transliteration scheme for English to Punjabi (2013)
15. To download. Guj-Ind-StyleGuide. http://download.microsoft.com/download/7/2/0/720b015e-94f9-4b6e-911f-539f38c60774/guj-ind-styleguide.pdf
16. Tithi (Internet). https://en.wikipedia.org/wiki/Tithi
17. Indian Place Names (Internet). http://www.irfca.org/docs/place-names.html
18. Gujarati Number names for Digits (Internet). https://www.omniglot.com/language/numbers/gujarati.htm