IJITKM 0973-4414 Volume-10, Number-2 Jan-June 2017 pp. 141-144 available online at www.csjournals.com

Named Entity Recognition for Dogri using ML

Shubhnandan S. Jamwal
PG Department of Computer Science and IT, University of Jammu

Abstract: Dogri language is a very low re-sourced language and Named Entity Recognition (NER) System aims to extract the existing information of the Nouns which can exists in the following categories such as: Name of the person, Organization, Location, Date, Time and Designation etc. Identifying a named entity remains a challenging task in all the Indian languages and it is considered as an important aspect for many natural languages processing (NLP) tasks. Much work on other Indian languages has been done but the Dogri language remains as a low resourced language. This paper explains the problem of NER in the framework of Dogri Language and explains the process of identification of the NE using machine learning approaches. For the training and testing, 950 words were used and we calculated performance accuracy of 72.51%.
Keyword: Machine learning, Dogri, Named Entity Recognition, Corpora.

INTRODUCTION

Dogri is an Indo-Aryan language and is a mother tongue of 422 million people. It is the second prominent language of J&K State, presence of Dogri language can also be felt in northern Punjab, Himachal Pradesh and other places. Dogri remained as a low resourced language in the field of machine transliteration.

Natural language processing is a field of Artificial Intelligence that deals with the methods of communicating with computers in natural languages like English, Hindi, Dogri etc. The researchers are developing number of NLP techniques for the different Indian languages for computational and analyzing processes which can enable a computer to understand the language. Number of different models of transliteration is available i.e., Grapheme-based Transliteration, Phoneme-based Transliteration, Hybrid-based Transliteration and Correspondence-based. There is a direct orthographical mapping from source graphemes to target graphemes, the transliteration key is pronunciation or the source phoneme rather than source grapheme, Hybrid-based simply combines grapheme and phoneme through linear interpolation and the fourth model can combine any number of grapheme or phoneme based models respectively.

In natural language processing the NER refers to the identification of proper nouns from natural language text and after the identification the individual units are categorized into different types such as person, location, date, time and organization. The applications of the NER are wide spread, therefore, numerous NER techniques and datasets are defined for the English and Indian languages.

Literature Review:

Jenny Rose Finkel and others[1] worked on nested named entities and they observed that many named entities contain other named entities inside them. They presented a new technique for recognizing nested named entities, by using a discriminative constituency parser. To train the model, they transformed each sentence into a tree, with constituents for each named entity. They worked on newspaper and biomedical corpora which contain nested named entities. Kriti Gupta [2] designed a system having two sub-tasks, first is identification and second is classification of the named entities. In first NER identifies words in texts which represent proper names like location, person-name, organization, date, time etc. and in second it classifies them in to predefined categories. Andrei Mikheev and others [3] claimed that extensive gazetteers lists of names of people, organizations, locations, and other named entities are sometimes considered as bottleneck in the design of Named Entity recognition systems. They report on a Named Entity recognition system which combines rule-based grammars with statistical (maximum entropy) models and reported on the system's performance with gazetteers of different types and different

sizes, using test material from the MUC-7 competition. They showed that, for the text type and task of this competition, it is sufficient to use relatively small gazetteers of well-known names, rather than large gazetteers of low-frequency names. Apurbalal Senapati and Arjun Das[4], described two systems for Named Entity Recognition (NER) and performance of two systems has been compared. The first system is a rule-based one whereas the second one is statistical (based on CRF) in nature. The systems vary in some other aspects too, for example, the first system works on untagged data (not even POS tag is done) to identify NER whereas the second system makes use of a POS tagger and a chunker. The rules used by the first system are mined from the training data. The CRF-based classification does not require any explicit linguistic rules but it uses a gazetteer built from Wiki and other sources. Erik F Tjong Kim Sang [5] applied a memory-based learner to the CoNLL-2002 shared task: language-independent named entity recognition and used three additional techniques for improving the base performance of the learner: cascading, feature selection and system combination. The overall system is trained with two types of features: words and substrings of words which are relevant for this particular task. The System is tested on the two language pairs that are Spanish and Dutch.

The experimental evaluation of the NER is conducted on Italian legal texts, and it is able to identify the classes of the ontology, as well as many hyponymy relations based on the approaches of Boella and colleagues [6, 7]. POS tagging can also be an effective technique complementing NER when calculating similarities [8,9,10]. In order to improve the NER in Portuguese language, this paper proposes a methodology for training text corpus based on Portuguese-language journalistic corpora. The Journalistic language has the best adherence to the contemporaneity of the language, since it preserves features such as objectivity, simplicity, impartiality, and is a reference of transmitting the information without ambiguity. The proposed methodology provides a model to extract entities and assess the obtained results with the use of Recurrent Neural Network architectures.

Noun in Dogri

The procedure for identification of nouns in Dogri language involves a variety of rules framed after morphological analysis of existence of nouns in Dogri language. The existence of noun in Dogri language resembles to a great extent with the existence of nouns in Hindi and Punjabi language. As illustrated in the algorithm above noun exists independently (Ram is good boy) or in combination with other grammatical categories that form a noun group. This noun group formation along with examples is shown in below table:-

Noun_group_formation	**Examples**
Noun	मेरा ना श्याम ऐ |
Noun-Na	मिग्गी बर्फ दिखना चंगा लगदा ऐ |
Mr./Mrs./Shri	श्री राम नाथ कोविंद जी ने उद्घाटन कित्ता |
Adjective + Noun	ओ चंगा जागत ऐ |
Noun+ch/ne/	मिग्गी श्याम ने दसया |
pronoun + Noun	ओ श्याम दे घर गया |
Quantifier + Noun	चार जागत स्कूल गए |
ne+Noun+gi	अध्यापक ने श्याम गी शाबाशी दित्ती |
Noun + Da	श्याम दा व्याह होया |
Noun + Di	श्याम ने राम दी कहानी जागतें गी सुनाई |

Noun Identification Algorithm:

```
1.  Procedure noun_identification (dataset, tokenization, stemmer, extraction)
2.  Variables:
3.      Txfl: Textfile.txt
4.      Snt ← extraction(Txfl)
5.      Morph ← tokenization(Snt)
6.      Morphstm ← stemmer(Morph)
7.      Resultvrb : String
8.  Begin:
9.     Snt [j] ← extraction(Txfl)
10.    While Snt[j] ≠ EOF do
11.       Morph[i] ← tokenization(Snt[j])
12.       While i < Morph.len do
13.          Morphstm[k] ← stemmer(Morph[i])
14.          i ← i + 1
15.        k ← k + 1
16.       While k < Morphstm.len
17.        If dataset.Noun_N(Morphstm[k]) = True then
18.           Display(Noun)
19.        else
20.           update_paradigm(Snt[j], Morphstm[k])
21.    j ←  j + 1
22.    end:
```

Update_noun_paradigm:

```
1.  Procedure update_paradigm( sentence, morp_h)
2.  Variables:
3.     Adj_a ← dataset.Adjective()
4.     Pron_n ← dataset.Pronoun()
5.     Sal_s ← Paradigm.getvalue(Salutation_s)
6.     Quant_f ← Paradigm.getvalue(Quantifiers_q)
7.     Ptn1 ← morp_h.Regex_Suffix( ना )
8.     Ptn2 ← morp_h.Regex_Succ( ने, च, गी, Adj_a)
9.     Ptn3 ← morp_h.Regex_Succ( दी, दा )
10.    Ptn4 ← morp_h.Regex_Pred(Quant_f , Adj_a)
11. Begin:
12.       If sentence.Pattern(Ptn1)= True || sentence.Pattern(Ptn2)= True || sentence.Pattern(Ptn4)= True then
13.         Display(morp_h + "=Noun")
14.         Paradigm_noun.add(Morph_h)
15.       Else if morp_h.pred(Sal_s) then
16.         Display(morp_h + "=Noun")
17.         Paradigm_noun.add(Morph_h)
18.       Else if morp_h.pred(Pron_n) then
19.         Display(morp_h + "=Noun")
20.
21.       Else if sentence.Pattern(Ptn1)= True then
22.             If dataset.identifyverb(morp_h) = True then
23.                Break;
24.             Else
25.               Display(morp_h + "=Noun")
26.               Paradigm_noun.add(Morph_h)
27.        Else
28.           Break;
29.   end:
```

Results and Discussions

After the completion of the initial processing of test data i.e., after extraction, tokenization and stemming, we have employed above algorithm for noun identification. We have performed four test runs of 9856, 13432, 18456, and 26578 morphs and got accuracy of 70.45%, 74.63%, 72.92% and 71.49%.

Test Case	Total no. of words	Accurately identified Nouns	Inaccurately identified Nouns	Accuracy (%)
Test-Run-1	9856	695	291	70.45
Test-Run-2	13432	1017	326	74.63
Test-Run-3	18456	1823	677	72.92
Test-Run-4	26578	2235	891	71.49
Average	**17081**	**1443**	**547**	**72.51**

We have attained an average accuracy of 72.51%.There are postpositions and prepositions used in Dogri language that makes the noun identification a challenging task because of the fact that these postpositions and prepositions are not uniquely exercised. We have a limited data set of pos tagged data involving Nouns, Pronoun and Adjectives, which is relevant to our work. Besides these we have maintained lists of Salutations used before Nouns and Quantifies that quantify nouns (two boys, second boy etc). We have made an attempt to capture some of the pre/post positions that are used with noun employing 3-gram approach. This approach suits well with noun identification as it covers the preceding and succeeding morph of nouns. We will continue our effort in enhancing the pos tagged dataset and morphological analysis of nouns utilization in Dogri language so that accuracy in identifying nouns can be improved.

References:

[1] Jenny Rose Finkel, Chris Manning and Christopher D. Manning, Nested named entity recognition, EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Volume 1, August 2009, Pages 141–150.

[2] Kriti Gupta, Named entity recognition: case study, ICWET '11: Proceedings of the International Conference & Workshop on Emerging Trends in Technology, February 2011, Pages 1357.

[3] Andrei Mikheev and Marc Moens and Claire Grover, Named Entity recognition without gazetteers, EACL '99: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, June 1999, Pages 1–8.

[4] Apurbalal Senapati and Arjun Das, Named-Entity Recognition in Bengali, FIRE '12 & '13: Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation, December 2013, Article No.: 14, Pages 1–5.

[5] Erik F Tjong Kim Sang, Memory-based named entity recognition, COLING-02: proceedings of the 6th conference on Natural language learning, Volume 20, August 2002, Pages 1–4.

[6] Peter McCullagh and John A. Nelder. 1989. Generalized Linear Models, (Chapman & Hall/CRC Monographs on Statistics & Applied Probability) (2 ed.). Chapman and Hall/CRC.

[7] Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. Semantic Processing of Legal Texts. Number 6036 in Lecture Notes in AI. Springer.

[8] X. Li and W. Croft. Novelty detection based on sentence level pattern. In CIKM'05, 2005.

[9] X. Li and W. B. Croft. An information-pattern-based approach to novelty detection. Information Processing and Management, vol.44: pp.1159–1188, 2008.

[10] K. W. Ng, F. S. Tsai, L. Chen, and K. C. Goh. Novelty detection for text documents using named entity recognition. In Information, Communications and Signal Processing, 2007 6th International Conference on, pages 1–5. 1, December 2007.