A Survey ON Named Entity Recognition for Indian Languages

Shalaka Naik Dessai
Department of Information Technology
Goa Engineering College
Ponda, Goa

Abstract— **Named Entity Recognition (NER) is a tool based on principles of Artificial Intelligence (AI) and Natural Language Processing (NLP) for automatically tagging Named Entities from unstructured text. Process of identifying and classifying all proper nouns into pre-defined classes such as persons, locations, organization, and others. Some of the application areas of NER include- Information Extraction and Information Retrieval, Machine Translation, Text Summarization, etc. Working in NER in Indian languages is a challenging task and limited due to a lack of resources. In this paper, we present a brief overview of NER for Indian languages.**

Keywords— Named Entity Recognition, Indian languages, Natural language processing, Classification

I. INTRODUCTION

Named Entity Recognition (NER) is a tool based on principles of Artificial Intelligence (AI) and Natural Language Processing (NLP) for automatically tagging Named Entities from unstructured text. Process of identifying and classifying all proper nouns into pre-defined classes such as persons, locations, organization, and others. Some of the application areas of NER include- Information Extraction and Information Retrieval, Machine Translation, Text Summarization, etc. Working in NER in Indian languages is a challenging task and limited due to a lack of resources. In this paper, we present a brief overview of NER for Indian languages.
Language is the necessary entity for human communication. To make the machines understand such kinds of natural languages, Natural Language Processing (NLP) is used. There has been growing interest in this field of research since the early 1990s. Named Entity Recognition is a two-step process: Identification and Classification. In the identification stage, proper nouns or NEs are identified. And those NEs are separated in their classes using classification.

PROBLEM FACED IN INDIAN LANGUAGES:
While significant work has been done in English NER, with a good level of accuracy, work in INDIAN LANGUAGES has started to appear only very recently. Some issues faced in Indian languages
1)Indian languages are relatively free-order languages.
2)There is no concept of capitalization of leading characters of names in Indian Languages unlike English and other European languages which play an important role in identifying Named Entities.
3) Some of the Indian languages like Assamese, and Telugu are agglutinative.
4)Unavailability of resources such as Parts of speech (POS) tagger, good morphological analyzer, etc for ILs. Name lists are found available on the web which is in English but no such lists for Indian Languages can be seen.
5) Indian languages are morphologically rich and highly inflectional.

There is a lot of work done in NER when it comes to English and other foreign languages. However, when we consider Indian languages particularly regional ones, we find that there is not much work done in them.

In this paper, we present a brief overview of NER for Indian languages.
Some languages like Hindi, Bengali, etc., have more work done as compared to other regional languages.

2.NER FOR HINDI

Kaur developed NER for Hindi using rule based approach which requires a set of rules which are manually written by linguistics. In this approach a huge gazetteer list is constructed for every named entity class. Rule based NER systems are considered as highly accurate systems as these systems require extensive knowledge of a particular language and domain to design syntactic-lexical pattern based rules .Their system identified three new named entities that were money value, direction values and animal/bird entities. Their system worked on new rule that is “no name entity rule” which had made improvement in existing rules. In list look up approach, different tables were created in database and named entities were extracted from these tables. The accuracy of their system was 95.77%.[1]

Saha et.al did the development of NER for Hindi using Machine Learning approach. Training data consists of 234 thousand words, collected from the newspaper “Dainik Jagaran” and is manually tagged with 17 classes including one class for not name and consists of about 16,482 Named Entities. This paper also reports development of a module for semi-automatic learning of the context pattern. The system was evaluated using a blind test corpus of 25K words having 4 classes and achieved an F-measure of 81.52%.[2]

Goyal focuses on building a NER for Hindi using CRF. This method was evaluated on test set1 and test set 2 and attains a maximum F1-measure around 49.2% and nested F1-measure around 50.1% for test set1 maximum F1-measure around 44.97% and nested F1-measure around 43.70% for test set2 and F-measure of 58.85% on development set. [3]

Sujan Kumar Saha developed a MaxEnt based NER Model for Hindi language in which he used many different features such as orthographic features (decimal, digits), affixes, left and right context words, part-of speech feature etc. For performance reasons, he used 8 gazetteer lists such as weekdays, organization end word lists, month's name , person prefix words list, location names list, first names list, middle names list and surnames list. The performance of the system was then evaluated against the blind test, set having the 4 classes - person, organization, location and date. This system achieved f-measure of 81.52%.[4]

Gupta and Arora describe the observation made from the experiment conducted on CRF model for developing Hindi NER. It shows some features which makes the development of NER system complex. It also describes the different approaches for NER. The data used for the training of the model was taken from Tourism domain and it is manually tagged in IOB format.[5]

Athavale, V., Bhardwaj, developed an end to end neural model based on bidirectional Long Short Term Memory (Bi-LSTM) for Hindi NER .Authors designed their model in two stages. In first stage, authors used the unlabelled corpus to learn word embedding based on skip gram approach and glove approach. In second stage, their system used bidirectional LSTM. System's embedding layers were initialized with learned word vectors for every word and then, system was trained end-to-end on labelled data. Their model achieved 77.48% accuracy for Hindi.[6]

3.NER FOR BENGALI

Asif Ekbal and Sivaji Bandyopadhyay proposed themethod that is the combination of Conditional Random Field (CRF) and Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) for NER in Bengali. They took about 272k word forms of training set for testing. And they developed the semi-supervised learning technique that uses the un-labeled data during the training of the system. They described that the use of large corpora is not enough but the system should measure to automatically select the effective documents and the sentences from the unlabeled data. The weighted voting approach is used to combine the models and the average experimental result of recall, precision, and f-score values is 93.79%, 91.34%, and 92.55% respectively. [7]

Hasan et.al presented a learning-based named entity recognizer for Bengali that do not rely on manually constructed gazetteers in which they developed two architectures for the NER system. The corpus consisting of 77942 words is tagged with one of 26 tags in the tagset defined by IIT Hyderabad where they used CRF++ to train the POS tagging model. Evaluation results shows that the recognizer achieved an improvement of 7.5% in F-measure over a baseline recognizer.[8]

Chaudhuri and Bhattacharya has made an experiment on automatic detection of Named Entities in Bangla. Three-stage approach has been used namely dictionary based for named entity, rules for named entity and left-right co-occurrences statistics. Corpus of Anandabazar Patrika has been used from the year 2001-2004. The manual tagging was done by the linguistic based on the global knowledge. Experimental results has shown the average recall, precision and f-measure to be 85.50%,94.24% and 89.51%.[9]

Mah Dian Drovo, Moithri Chowdhury, Saiful Islam Uday, Amit Kumar Das developed a method which is using both ML and Rule Base approach together for NER based on Bengali language. Mainly the rule based approach has been merged with ML. For ML Hidden Markov Model (HMM) and for rule base approach Regular Expression has been used. A Named Entity (NE) tagged corpus has been developed by using Bengali newspaper, which consists of 10k words that has been manually annotated with seven tags. This paper concludes with experimental results which shows two distinctive ways of our proposed model.[10]

4.NER FOR MALAYALAM

Ajees A Pa developed a NER system for Malayalam using neural networks. The proposed system utilizes different features such as POS information of the word, embedded representation of words and suffixes, POS information of preceding words, etc. A corpus of 20615 sentences was used for training and testing. Out of vocabulary words were handled using the online training feature of Word2Vec. The overall accuracy in NER of the proposed system is 95.3%.[11]

5.NER FOR ORIYA

Biswas et.al presented a hybrid system for Oriya NER that applies both ME and HMM and some handcrafted rules to recognize NEs. Firstly the ME model is used to identify the named entities from the corpus and then this tagged corpus is regarded as training data for HMM which is used for the final tagging. Different features have been considered and linguistic rules help a lot for identification of named entities. The annotated data used in the system is in IOB format. Finally the system comes with an F-measure between 75% to 90%.[12]

6.NER FOR KONKANI

Annie Rajan and Ambuja Salgaonkar developed NER using Konkani language. Gold data of 1000 NER-tagged Konkani sentences consisting of 1068 named entities is one of the linguistic resources generated through this work. A conditional random field (CRF) classifier built on the training data set of 794 named entities from 800 sentences of the corpus, demonstrated 96% accuracy and 72% f-score. On the test data set of 274 named entities from 200 sentences of the corpus, 86% accuracy and 66% f-score were obtained. When the training and test data were complemented with a lookup table consisting of a database of 12 months, 53 locations, 44 person-names and 23 numerals and their synonyms, the figures improved to 99% accuracy and 90% f-score for the training data set, and 89% accuracy and 73% f-score for the test data set.[13]

7.NER FOR TAMIL

Vijay Krishna and Sobha developed a domain-specific Tamil NER for tourism by using CRF. It handles morphological inflection and nested tagging of named entities with a hierarchical tag set consisting of 106 tags. A corpus of 94k is manually tagged for POS, NP chunking, and NE annotations. The corpus is divided into training data and the test data where CRF is trained with the former one and CRF models for each of the levels in the hierarchy are obtained. The system comes out with an F-measure of 80.44%.[14]

Pandian et. Al presented a hybrid three-stage approach for Tamil NER. The E-M(HMM) algorithm is used to identify the best sequence for the first two phases and then modified to resolve the free-word order problem. Both NER tags and POS tags are used as the hidden variables in the algorithm. Finally, the system comes out with an F-measure of about 72.72% for various entity types.[15]

Srinivasagan, Jeyashenbagavalli and Suganthi developed a NER model for the Tamil Language, usinga Hybrid approach that uses both Rule-Based and Hidden Markov Model in succession, which identifies the person, location and organization names. The system identifies unknown entities using statistical Hidden Markov Model. This Named Entity Recognizer shows that the Hybrid approach is better than the statistical model of HMM with overall Precision, Recall and F-Score values as 93.18%, 86.54% and 89.7% respectively through only a limited sized corpus.[16]

8.NER FOR MARATHI

Nita Patil developed a NER system for Marathi language that applies Hidden Markov Model, language specific rules and gazetteers to the task of named entity recognition (NER) in Marathi language. Starting with named entity (NE) annotated corpora and lemmatization first a baseline NER system was implemented. Then some language specific rules are added to the system to recognize some specific NE classes. Also, some gazetteers and context patterns are added to the system to increase the performance. After preparing the one-level NER system, a set of rules are applied to identify the nested entities. The system was able to recognize 12 classes of NEs with 89.05% accuracy in average NE identification and 90.09% accuracy in average NE classification for held out and unseen test datasets in Marathi.[17]

Patawar and Potey developed the NER for Marathi language tweets. In their system, Patawar et al. have used a hybrid of k-Nearest Neighbour and CRF. Initially, the normalized tweets are assigned the confidence value 'cf ' using a K -value of 4. Then the CRF labeler is used to assign a label. The assigned token is added to the clusters and the system uses it for the further training. Otherwise, the CRF label is assigned to it. Only Location and Names are identified by them with a precision of 39.80 and recall of 85.11 for location and a precision of 59.72 and recall of 25.28 for the name tags. CRF makes it easier to add the prefixes and suffixes which is necessary for NER in Indian languages.[18]

9.NER FOR TELGU

CRF calculates the probability, however if 'cf' exceeds then Srikanth and Murthy used part of LERC-UoH Telugu corpus where CRF based Noun Tagger built with 13,425 words manually marked data and tested in a test set of 6,223 words and came out with F-average 91.95%. Then they create a NER based on the law a program with 72,152 words including 6,268 Named Businesses where they identified specific issues related to Telegu. The NER also later developed a CRF-based NER system for Telegu and obtain a total F rating of between 80% and 97% in various tests.[19]

Shishtla et.al conducted an experiment on the development data released as a part of NER for South and South East Asian Languages (NERSSEAL) Competition. The Corpus consisting of 64026 tokens was tagged using the IOB format (Ramshaw and Marcus, 1995). The author have showed experiments with various features for Telugu. The best performing model gave an F-1 measure of 44.91%.[20]

Raju et.al have developed a Telugu NER system by using ME approach. The corpus was collected from the iinaaDu, vaarta news papers and Telugu Wikipedia. Manually tagged test data is prepared to evaluate the system. The system makes use of the different contextual information of the words and Gazetteer list was also prepared manually or semi-automatically from the corpus and came out with a an F-measure of 72.07% for person, 6.76%, 68.40% and 45.28% for organization, location and others respectively.[21]

CONCLUSION:

In conclusion, the study of Named entity recognition was done for several Indian languages. Some languages such as Bengali and Hindi had comparatively extensive work done in this area. On the contrary, languages such as Telugu and Marathi had less work done in the field of named entity recognition. Languages that already had sufficiently accurate processing tools required much less investment in terms of time and money as compared to those languages, which had primitive language processing tools.

REFERENCES:

[1] Kaur, Y. and Kaur, R. "Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach", In IJSRM, vol. 3(3), pp: 2300-2306, 2015.

[2] S. K. Saha, S. Sarkar, and P. Mitra, "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition," in Proceedings of the 3rd International Joint Conference on NLP, Hyderabad,India, January 2008, pp. 343–349.

[3] A. Goyal, "Named Entity Recognition for South Asian Languages," in Proceedings of the IJCNLP-08 Workshop on NER for South and SouthEast Asian Languages, Hyderabad, India, Jan 2008, pp. 89–96.

[4] Saha, Sujan Kumar, Sudeshna Sarkar, and Pabitra Mitra. "A hybrid feature set based maximum entropy Hindi named entity recognition." Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I. 2008.

[5] P. K. Gupta and S. Arora, "An Approach for Named Entity Recognition System for Hindi: An Experimental Study," in Proceedings of ASCNT2009, CDAC, Noida, India, pp. 103–108.

[6] Athavale, V., Bhardwaj, S., Pamecha, M., Prabhu, A. Shrivastava, S., "Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Scarcity," In NLPAI, 2016, pp: 154-160.

[7] Ekbal, Asif, and Sivaji Bandyopadhyay. "Named entity recognition using appropriate unlabeled data, post-processing and voting." Informatica 34.1 (2010).

[8] K. S. Hasan, M. ur Rahman, and V. Ng, "Learning -Based Named Entity Recognition for Morphologically-Rich Resource-Scare Languages," in Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 2009, pp. 354–362.

[9] B. B. Chaudhuri and S. Bhattacharya, "An Experiment on Automatic Detection of Named Entities in Bangla," in Proceedings of the IJCNLP08 Workshop on NER for South and South East Asian laanguages, Hyderabad, India, January 2008, pp. 75–82.

[10] M. D. Drovo, M. Chowdhury, S. I. Uday and A. K. Das, "Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach," 2019 7th International Conference on Smart Computing & Communications (ICSCC), 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843661.

[11] Ajees, A. P., and Sumam Mary Idicula. "A named entity recognition system for Malayalam using neural networks." Procedia computer science 143 (2018): 962-969.

[12] S.Biswas, S.P.Mohanty, S.Acharya, and S.Mohanty, "A Hybrid Oriya Named Entity Recogntion system," in Proceedings of the CoNLL, Edmonton, Canada, 2003.

[13] Rajan A, Salgaonkar A. Named Entity Recognizer for Konkani Text. InICT with Intelligent Applications 2022 (pp. 687-702). Springer, Singapore.

[14] V. R and S. L, "Domain focussed Named Entity Recognizer for Tamil using Conditional Random Fields," in Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, Hyderabad, India, 2008, pp. 59–66.

[15] S. Pandian, K. A. Pavithra, and T. Geetha, "Hybrid Three-stage Named Entity Recognizer for Tamil," INFOS2008, March 2008.

[16] K. G. Srinivasagan, S. Suganthi and N. Jeyashenbagavalli, "An *Automated System for Tamil Named Entity Recognition Using Hybrid Approach," 2014 International Conference on Intelligent Computing Applications, Coimbatore, 2014, pp. 435-439.*

[17] Patil, Nita, Ajay Patil, and B. V. Pawar. "Hybrid Approach for Marathi Named Entity Recognition." Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). 2017.

[18] M. L. Patawar and M. A. Potey, "Extending hybrid Conditional Random *Fields approach of Named Entity Recognition for Marathi tweets," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, 2016, pp. 1-5.*

[19] P.Srikanth and K. N. Murthy, "Named Entity Recognition for Telegu," in Proceedings of the IJCNLP-08 Wokshop on NER for South and South East Asian languages, Hyderabad, India, Jan 2008, pp. 41–50.

[20] P. M. Shishtla, K. Gali, P. Pingali, and V. Varma, "Experiments in Telegu NER: A Conditional Random Field Approach," in Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian laanguages, Hyderabad, India, January 2008, pp. 105–110.

[21] G. Raju, B.Srinivasu, D. S. V. Raju, and K. Kumar, "Named Entity Recognition for Telegu using Maximum Entropy Model," Journal of Theoretical and Applied Information Technology, vol. 3, pp. 125–130, 2010.