JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A survey on Multingual name entity recognition using Natural language processing

Sanjay Kumar Duppati, Dr.A.Ramesh Babu

Research Scholar, Professor

Chaitanya Deemed to be University ,Hanmakonda, Telangana.

Abstract: Named Entity Recognition (NER) is a key component in NLP systems for question answering, information retrieval, relation extraction, etc. NER systems have been studied and developed widely for decades. Machine learning and Statistical techniques are powerful analysis tools yet to be incorporated in the new multidisciplinary field diversely termed as natural language processing (NLP) or computational linguistic. The linguistic knowledge may be ambiguous or contains ambiguity; therefore, various NLP tasks are carried out in order to resolve the ambiguity in speech and language processing. In this paper, we explore various methods that are applied to solve NER and a survey is done on various approaches used to recognize name entity in various Indian languages.

Keywords: Named Entity Recognition, natural language processing, Machine translate.

I. INTRODUCTION

Natural language processing (NLP) is a subfield of artificial intelligence and computational linguistics that studies the problems of the machine age and the experience of herbal human languages. In specific NLP, it presents with reading, knowledge and language technology that humans obviously use in the way humans interact with computers in every written and spoken context, and using herbal human languages instead of human languages. A fully NLP-based machine aims to design and build a program that can also provide complete and unique additional registers in response to the actual information the consumer wants. NLP consists of the processes: natural language understanding (NLU) and natural language generation (NLG). The NLU structures transform human language samples into more formal representations along with wood analysis or first-order common sense structures that are less complex to control for computer applications.

Named Entity Recognition:

The task of processing the text to identify the classes of names such as names of person, places, organizations etc.

Kumar and Bhattacharyya [2006] stated that NER is an data mining sub-task that seeks to discover and categorize atomic factors within textual content into pre-defined categories that include names of people, organizations, places, expressions of times, quantities, monetary values, probabilities, and many more. Also called entity identity and entity extraction. Therefore, the popularity of named entities is the task of identifying and classifying all proper names in a textual content. It is a step-by-step identification and class of named entities.

The categories chosen for a specific NER task may depend on the company's requirements, for example, if the geographical category is dynamic, it may be necessary to classify each place feature as a specific form of location including CITY, STATE, COUNTRY, etc. Numerical classification is critical to a specific region, and then the types that deal with digital records may also need to be more precise. In NER on the basis that one tag can be assigned to a phrase in context, so the NER version will assign each one of the required masterpieces or a NOT-aNAME tag to stand for "none of the preferred phrases". We prepare states (named entity data) in regions, one place for each preferred category plus one place for NOT-ANAME. In addition, there are two unique cases, the beginning of the sentence and the end of the sentence. An arbitrary variety of classes can be provided to the system at runtime.

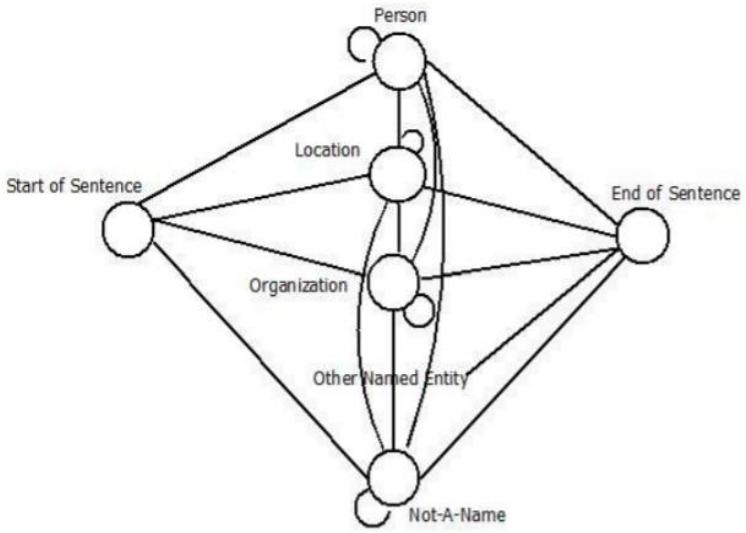

Fig.1 elucidates this concept diagrammatically

MOTIVATION

NER has become a primary research area in NLP due to its sensitive programs in various fields. However, it requires modern technical innovations, which must be more accurate and efficient than current strategies. All Indian languages use SOV (Subject Object Verb) word order, while English use SVO (Subject verb object) word order, and due to loss of capitalization, fonts, standardization, and spelling variation, the English NER techniques cannot be used at the same time for Indian languages.

In previous work done in the field of NER within the Hindi language, researchers have shown that it is very difficult to design an entirely rule-based machine for Indian languages, and due to the limited availability of applicable resources in Indian languages, it is very difficult to obtain an accurate NER statistical machine. Therefore, there is a need for a unique and accurate Indian NER device expansion. The artwork inside the NER English language machine results in an almost human performance. But in the Hindi context, not much has been done on NER for Indian languages and the accuracy level of Indian NER is equally lower compared to English. Gift studies come close to achieving accuracy in the NER of Hindi, guiding humans as far away as possible. Since all Indic languages have remarkable similarity in structure, most of them use the Sanskrit-derived Devanagari alphabet and use SOV (subject object verb) sentence order by default, so there is an opportunity to design a unified model for all Indic languages. In this sense, the current research is inspired by the design of an unusual version of NER for all Indian languages. Also to augment a version, which can be used for NER for another language, let's say below, with a touch mode.

As we realize ourselves, the status of NER in Indian languages depends on the degree of research and there is no work in the product stage. Given this, the current research aims to extend a software program that can do NER in any Indian language with little or no human effort.

Since evaluating the accuracy of program output is very difficult and requires many hours of work and efforts, this looks at the goals to design a solution for automated evaluation of NER output.

APPLICATIONS OF NER

There are a number of applications where Named Entity Recognition is required such as

- Machine Translation,
- Intelligent document access,
- Cross Lingual Information Retrieval (CLIR),
- Summarization,
- Question Answering System

II. LITERATURE SURVEY

Wang et al. [2020] First, they presented the preschool structure and tasks of 4 unusual preschool models: BERT, ERNIE, ERNIE2, Zero-tiny and RoBERTa. They then implemented these NER pre-training models using an appropriate level 1 tier, and compared the consequences of the unique architecture and preschool assignments on the NER mission. Test results showed that RoBERTa conducted the most recent results in MSRA-2006 Named Entity Recognition (NER) which was a core Natural Language Processing (NLP) project to extract entities from unstructured statistics. Previous NER technologies were based on knowledge

acquisition or an in-depth study of the machine. Recently, pre-educational innovations have made significant progress in performing multitasking NLP.

Amrita Anandika et al. [2019] the main genre focused on understanding the different types of NER and the tactics applied for NER, specifically the different patterns of machine learning used to identify named entities. Natural language processing (NLP) has become a sub-component of artificial intelligence, which typically specializes in training computers to understand and work with human languages and brings computers closer to understanding language like lo makes a human. Named entity recognition (NER) has become the core of NLP systems. NER is a computerized identification procedure for specific entities in a specific text or report content. The named entities were either real-world objects or in the named standard entities, they were proper names like individual name, place, date, time, etc. Research areas such as question answering, summary systems, information mining, machine learning, bioinformatics, semantic web search, video annotations, and much more.

Ji Young Lee et al. [2018] Recent processes based on artificial neural networks (ANN) have shown promising results for the recognition of entities called (NER). To get excessive returns, you must study NNAs on a large set of labeled data. However, it may be difficult to obtain the labels for the dataset on which the user wishes to perform the NER: the lack of labels is indicated specifically for the affected person. Quantitative change analysis can identify this problem as well. Specifically, they show that transferring an ANN model learned on a large, labelled data set to another data set with a limited number of labels improves contemporary outcomes on two unique datasets so that the patient is familiar with identification.

Zhao et al. [2017] According to the text statistics of the Chinese traditional medicine weight loss program, thinking about the characteristic of the Chinese medicine entity's reputation, the statistics of the small group are generated, and the entity's reputation experience with the information from the self-generated group. In this article, we use a conditional random area approach that relies entirely on the device study model to train and monitor unusual medical illnesses: anemia, cough, hypertension, hyperlipidemia, and diabetes. The target entity is extracted within the medical aspects of the framework elements and the dialectical scientific aspects of two categories. By including the structural functions of sentences and expanding group records, experimental effects show that the proposed method has better accuracy and better memory load, and the correctness of adding structural features is confirmed.

Chandranath Adak et al. [2016] an approach has been proposed to identify Named Entities (NE) directly from images of unstructured handwritten offline reports with no apparent popularity of individuals/words, and with very few useful resources from text and language based regulations. In the pre-processing stage, the report image changed to binary, after which the text content changed to word segmented. Sentence/skew/baseline deviation corrections were also performed. After pre-processing, the statements were submitted for the NE reputation. They analyzed the structural and objective features of boundary suspicions and extracted some applicable features from the picture of the sentence. Then the neural community BLS™ became used to learn about NE. Their devices also have a post level processing level to reduce true NE rejection fees. The proposed technique produces encouraging results in the images of every

historical and recent report, comprised of those in the Australian Archives, which can be viewed here for the first time.

Dong Wang et al. [2015] Transfer study is a vital way to generalize trained models for one preparation or project to other settings or commitments. For example, in speech hearing, a skilled phonemic model of one language can be used to understand speech in any other, with very little data on re-education. Acquisition of knowledge transfer is closely related to multitasking learning (navigation language vs. multilingual), and has historically been studied under the name "version modification". The recent increase in deep study indicates that the field of changes becomes simpler and more efficient with high-level abstract functions that are learned through the use of deep models, and 'transformation' may be more easily achieved not only between distributions of records and fact classes, but also between version systems (e.g.,, surface networks and deep networks) or possibly version types (e.g., Bayesian fashion, neural fashion). This evaluation document summarizes some of the current notable studies toward this pathway, particularly in relation to speech and language processing. In addition, we present some findings from our group and highlight the potential of this exciting research discipline.

Alex Graves et al. [2014] This paper provided a speech recognition device that instantly transcribed audio information with text, without the need for an intermediate audio representation. The machine became a combination of the LSTM bidirectional recurrent neural network architecture and the objective function of temporal communicative classification. A modification to the target feature was delivered that enables the community to reduce the expectation of the arbitrary text loss feature. This allows for immediate improvement of the phrase error rate, even in the absence of a lexicon or language version. The device ended up with a phrase error rate of 27.3% in the Wall Street Journal group and without using previous language data, 21.9% with the simplest lexicon of allowed words and 8.2% in the Trigram language version. Integrating the community with a referral device reduces the error rate to 6.7%.

Graves et al. [2013] recurrent neural networks (RNN) were a powerful version of sequential statistics. Extensive training methods that include link temporal classification make it possible to train RNNs on serial tagging problems in which the input-output alignment is unknown. The combination of these techniques with the short-term memory RNN architecture has proven fundamentally fruitful, providing recent influences on the popularity of cursive writing. However, the overall performance of RNN in speech reputation has been disappointing thus far, with the best results obtained through deep feed networks. This article conclusively examined deep recurrent neural networks, which combined more than one stage of representation that have been shown to be very effective in deep networks with the flexible use of the diverse contexts that enable RNNs. When trained in smoking cessation with appropriate regulation, they determined that the long and short memory RNNs scored a test group error of 17.7% on the TIMIT audio popularity scale, which according to our information is the logged first-level rating.

Mohammed et al. [2012] This system is designed based on neural network technology. An important task of neural society technology is to automatically learn to capture patterns of objects and make smart selections based primarily on available records, moreover, it can be implemented to classify new information within

huge databases. It is proposed to use machine learning technology for NER classification of Arabic textual content based entirely on neural network technology. Neural network technology has succeeded in many areas of artificial intelligence. The machine consists of three stages: the first level is the pre-processing that cleans up the collected facts, the second includes the conversion of Arabic letters into Roman alphabets and the final stage applies the neural community to classify the accumulated information.

Zheng et al. [2010] Web text category is a procedure for routinely selecting types of text content within a given category, according to the content of the text content. Web text classification system uses machine knowledge, experience engineering and related various fields of understanding, Internet access in the text, after text pre-processing, Chinese word segmentation and learning classifier, using the type algorithm to implement automatic class. This article designed a network version of the Chinese text classification machine and the system examined, the experimental results show that the Internet text classification machine has the main characteristics, which can be efficiency and accuracy.

Ekbal et al. [2009] a series of strategies are proposed for each classifier's output technology to reduce errors and further improve performance. Finally, we use 3 weighted voting strategies to mix male and female exhibitors. The experimental results showed the effectiveness of the proposed multi-engine method with the overall values of recovery, accuracy and F-score 93.98%, 90.63% and 92.28%, respectively, indicating a 14% improvement in the outcome. F on a well-performing SVM-based machine and an F-Score score 18.36% across the least popular platform. Benchmarking results also show that the proposed system is superior to the three different existing Bengali NER systems.

Zhang et al. [2008] This paper given the ability of a conditional random field (CRF) that combined with more than one function to perform strong and accurate identification of the named Chinese entity. Half of more than one feature template that includes local function templates and global feature templates used to extract two features with the help of human information. Besides, we show that human information can reasonably clean the release and then the need for training data for CRF is reduced. From the experimental results on the daily group of people, we will conclude that our model is an effective sample to combine the statistical model with human understanding. Experiments over any other facts also corroborate the above conclusion, which shows that our features have consistency in an experiment unique to the data.

III. VARIOUS APPROACHES FOR NER

Through a large review of the literature, we have come here to realize that there are two types of strategies used to understand the specific entity: rule-based learning and machine learning.

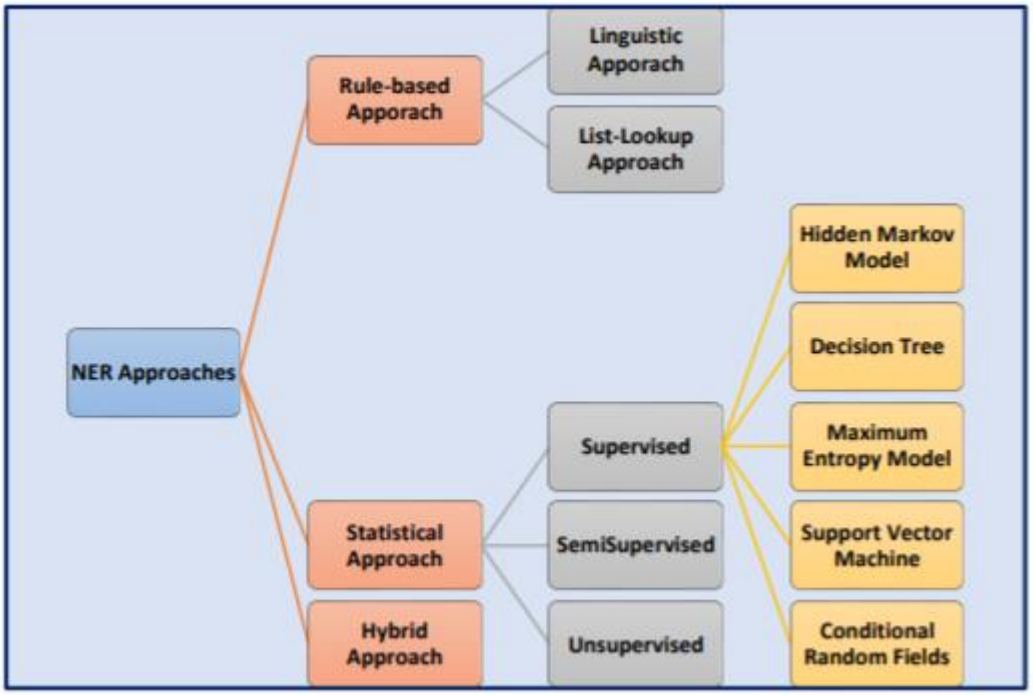

Fig.2 NER approaches

Rule-based approaches Lacks the ability to deal with durability and portability issues. Each new source of textual content requires a gigantic rebuilding of regulations to maintain peak overall performance, and renewal costs can be very high.

Machine-learning approach, is much more attractive, trainable and adaptable, maintaining a machine acquisition device is much cheaper than a primarily rule-based one, and provides basic statistical information not available to the human professionals, who developed the guidelines. However, Guodongzhou [2005] states that the performance of a machine-based machine is consistently weaker than that of a rule-based machine by about 2%. Machine study strategies consistently provide important statistical data that is not available to human specialists.

It is more attractive, trainable and adaptable, maintains a much cheaper machine acquisition device than a primarily rule-based one, and provides key statistical information not available to the human professionals who developed the guidelines. However, Guodongzhou [2005] states that the performance of a machine-based machine is consistently weaker than that of a rule-based machine by about 2%. Machine study strategies consistently provide important statistical data that is not available to human specialists.

Zeynel [2008], in the supervised method, pre-described class labels are assigned to documents based on the recommended probability using a learning set of labeled files. Also called learning with the teacher. Supervised learning is the area of mapping between input and output, and device output can be expected if new inputs are received. If the output takes a finite set of discrete values indicated by the input class labels, then the detected assignment ends up in the input facts class.

Since supervised mastery is simple and provides greater computational efficiency compared to unsupervised study strategies, it is assumed that the labels in school records are derived from the reality of the trusted floor,

which increases the amount of data. Captured within the release. Therefore, the supervised learning method was favoured by the learning community of the system.

IV. CHALLENGES FOR NER IN INDIAN LANGUAGES

Like NLP investigations in Indian languages, NER investigations in Indian languages also face different demanding situations some of them:

No capitalization – Indic languages lack large registers, which play an important role in identifying entities named in English.

Unavailability of large gazetteer – Internet resources for name lists (e.g., list of individual names, city names, etc.) are required as sources for NER. But these are only available to English-language entities and not available in Indian languages, which subsequently forces the use of transliteration or the creation of such dictionaries.

Lack of standardization and spelling - Indian individual nouns are huge and varied and many of these words can be found in the dictionary with specific meanings. Moreover, many of these names are also used as common names.

Inflectional language - India languages offer a rich and challenging set of linguistic and statistical properties that lead to long and complex word forms.

Scarcity of resources and tools - Indic languages are low-resource languages: annotated corpora, superior morphological parsers, POS tags, etc. Not yet available even brand grade.

V. CONCLUSION

The Named Entity Recognition (NER) field has been thriving for more than fifteen years. Its goal is to extract and categorize rigid assignment signs, from textual content, along with appropriate nouns and temporal expressions. In this survey, we established that the previous work ended in Indian languages. This survey was done in Indian languages, such as Telugu, Hindi, Bengali, Oriya and Urdu. A review of contracted strategies for developing NER systems, documenting current fashion away from the bases of artisans closest to the machine learning methods. Handcrafted structures provided precise overall performance for a relatively high system engineering cost.

FUTURE WORK

- The performance can further be increased by improving gazetteer lists.
- Analysing the performance using other methods like Maximum Entropy (ME) and Support Vector Machines (SVM)

- Comparing the results obtained by using various methods and evaluating the most accurate method for it.
- Improve the performance of every NE tag to make it overall more accurate.

REFERENCES

1. R. R. Salakhutdinov and N. Srivastava and, 2012, "Multimodal learning with deep boltzmann machines," pp. 2222–2230.
2. L. Zhang, and Z. Jin, 2015, "Distilling word embeddings: An encoding approach".
3. T. Mikolov and J. Dean, 2013, "Efficient estimation of word representations in vector space," arXiv preprint arXiv: 1301.3781, 2013.
4. T. Mikolov and I. Sutskever, 2013, "Exploiting similarities among languages for machine translation,"
5. C. Xing, D and Y. Lin, 2015, "Normalized word embedding and orthogonal transform for bilingual word translation,"
   Yu Wang Zuchang Ma, 2020, "Application of Pre-training Models in Named Entity Recognition", pp.23-26.
6. Amrita Anandika, Smita Prava Mishra, 2019, "A Study on Machine Learning Approaches for Named Entity Recognition", pp.2800311-2800322.
7. Ji Young Lee P. Szolovits, 2018, "Transfer Learning for Named-Entity Recognition with Neural Networks", 2018.
8. Y. Zhao, "Research on Entity Recognition in Traditional Chinese Medicine Diet", 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017.
9. Chandranath Adak, Michael Blumenstein, 2016, "Named Entity Recognition from Unstructured Handwritten Document Images", pp. 375- 380.
10. Dong Wang and Thomas Fang Zheng, 2015, "Transfer learning for speech and language processing", IEEE, pages 1225–1237.
11. Alex Graves, Navdeep Jaitly, 2014, "Towards End-to-End Speech Recognition with Recurrent Neural Networks".
12. A. Graves and G. Hinton, 2013, "Speech recognition with deep recurrent neural networks," IEEE, (ICASSP), 2013, pp. 6645–6649.
13. Mohammed, N.F. & Omar, N. (2012), "Arabic named entity recognition using artificial neural network", pp.1285- 1293.
14. Zheng, G. & Tian, Y. (2010, "Chinese web text classification system model based on Naive Bayes', ICEEE, pp. 1-4.
15. Ekbal, A., & Bandyopadhyay, S, 2009, "Named entity recognition in Bengali: A multi-engine approach". Proceeding of the Northern European Journal of Language Technology, pp. 26–58
16. Zhang, Y and Zhang, T, 2008, "Fusion of multiple features for chinese named entity recognition based on CRF model Information Retrieval Technology": Springer, pp. 95-106.