International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

A Probability based Classification of Named Entities for Malayalam Language combining Word, Part of Speech and Lexicalized features.

Gowri Prasad, Prakruthi S T

Abstract: ***Named Entity Recognition is the process wherein named entities which are designators of a sentence are identified. Designators of a sentence are domain specific. The proposed system identifies named entities in Malayalam language belonging to tourism domain which generally includes names of persons, places, organizations, dates etc. The system uses word, part of speech and lexicalized features to find the probability of a word belonging to a named entity category and to do the appropriate classification. Probability is calculated based on supervised machine learning using word and part of speech features present in a tagged training corpus and using certain rules applied based on lexicalized features.***

Index Terms: ***Named Entity Recognition, Named Entities, Lexicalized features, Supervised Machine Learning.***

I. INTRODUCTION

Named Entity Recognition (NER) does the process of finding the named entities present in a document using different approaches. Named entities are the designators of a sentence like person names, place names, institution names or other domain dependant names. NER is a very important application area of natural language processing. NER helps to make the tasks of information extraction much easier which in turn is highly useful for summarisation systems and other natural language processing applications. The approaches generally applied for named entity recognition are rule based approaches, machine learning based approaches and hybrid approaches. Rule based approaches are highly dependent on the language and the various resources available in that language. The resources include lists, word-nets etc. belonging to a language. Machine learning based approaches are again classified into unsupervised, supervised and semi-supervised approaches. Hybrid approach talks about combining two different approaches to form a new approach. Unsupervised and semi-supervised approaches does not require much human intervention as it employs the techniques of pattern recognition, automatic signature generation etc. While supervised machine learning approaches requires lot of human intervention and supervision to form an annotated corpus. The main challenge in developing NER systems is the ambiguous nature of the natural or spoken languages. Moreover unlike English, languages like Malayalam and other South Indian languages are mostly free word order in nature, lacks capitalization information and also lacks different resources like annotated training files, word lists etc. This makes the development of NER systems in such languages highly challenging. This paper proposes a probability based classification of named entities in Malayalam belonging to tourism domain. The proposed system makes use of supervised machine learning CRF based model for NER using word and part of speech features as proposed in [1]. Also, it combines some rules formed from the lexicalized features of the language to calculate the probability. Related works which has been carried out in the area of NER is briefly discussed in Section II, and section III explains the supervised machine learning approach used by the proposed model. The proposed named entity recognition system and its various stages are discussed in section IV. Section V includes the results and discussions followed by conclusion in section VI.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Gowri Prasad*, Information Science and Engineering, New Horizon College of Engineering, Bengaluru, India.
Prakruthi S.T, Information Science and Engineering, New Horizon College of Engineering, Bengaluru, India.

II. RELATED WORKS

The research work related to named entity recognition is progressing as it is highly used in many natural language processing applications. NER using Rule based approach applied to different languages is proposed in [2] and [3]. Unsupervised approaches for NER are described in [4] and [5] and the same unsupervised approach applied to specific domain is proposed in [6]. Languages which are rich in resources like tagged corpuses, word lists etc. uses more of supervised machine learning approaches. Hidden Markov Model based NER system is proposed in [7] and [8] and in [9] the author proposes a method of NER using decision trees. In [10] support vector machines concept is used in the process of NER. NER systems for Indian languages are also being developed. Conditional Random Field based NER for Manipuri language is proposed in [11] and the same for Tamil language is experimented in [12] . A hybrid approach for NER has been suggested in [13] that combines two approaches i.e. the rule-based approaches and machine learning based approaches. A comparative study of Conditional Random Field (CRF) approach and Maximum Entropy (MaxEnt) approach applied to English and Hindi language is described in [14].

Conditional Random Field approach for NER in Malayalam is proposed in [1] and various other similar NER models for the South Indian language viz. Malayalam is discussed in [15] and [16].

III. SUPERVISED MODEL

The supervised model used in this system is based on probabilistic Conditional Random Fields and that is a discriminative model. It assigns a probability for the possible labels based on the current, past and future observations or in other words the observation sequence. Given the sequence of current, past and future observations, CRF based models build a single exponential model that determines the joint probability of label sequence. CRF based models are globally conditioned on the observation sequence and hence it does not show local (based on each state) normalization of transition probabilities. It also does not create the problem of bias while labelling.

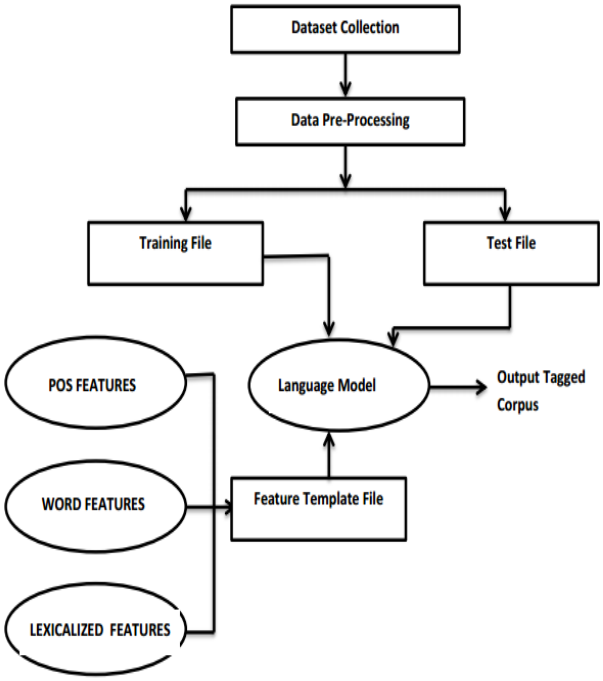

Fig.1. System Architecture of the Proposed Named Entity Recognition System

IV. THE PROPOSED METHODOLOGY FOR NAMED ENTITY RECOGNITION

The system discussed here uses the CRF based supervised machine learning model as proposed in [1] and also combine the rules formed from the lexicalized features available for the Malayalam Language. Fig 1 shows the system architecture of the proposed named entity recognition system.
The different stages:

- Data Gathering and Dataset Pre-processing.
- Formation of rules based on the lexicalized features.
- Preparation of the file with feature templates.
- Formation of the file for model training and testing.
- Training the model.
- Testing and Result Analysis.

A. Data Gathering and Dataset Pre-processing

Data gathering or Dataset collection is the process of collecting the raw data for training the model and data pre-processing is the process of making the raw data ready for training. The dataset used for this model consists of more than 1,00,000 words in Malayalam from tourism domain. After collection of the dataset the data was cleaned and tokenisation was performed. Tokenisation is the process of splitting the sentence into separate words. Appropriate part of speech (POS) tags were added to these words. POS tagging for the seed dataset was done manually and this seed dataset was given as training file for CRF model to learn and further perform POS tagging of the remaining words. Then chunk tags were added for each word manually.

B. Formation of rules based on Lexicalized features

Lexicalized features depend on the morphology of each language. Rules were framed by linguists of Malayalam language. These rules mainly dealt with prefix suffix associated with each word. Various classes of named entities mostly will have similar kind of prefixes or suffixes when used in different contexts. This relationship were identified and framed into rules which were added as feature for training. Also, rules were framed based on the surrounding word information. The probable words which can be joined or surrounded with different classes of named entity were recognized for rule formation using surrounding words.

C. Preparation of the file with feature templates

Next stage is the preparation of the feature template file. The features which were used in forming this model were word features, POS features and lexicalized features. Rules formed in the previous phase were used to encode the lexicalized feature in the feature template file. Unigrams and Bigrams were formed for encoding the POS features and the word features. Unigram feature considers only the previous and next word for the word feature and part of speech of the previous and next words for POS features. Bigrams consider the preceding and succeeding two words or part of speech of two words. In a similar way unigrams and bigrams corresponding to chunk tag features were also formed and added to the feature template file.

A few unigram features from the feature template file:

U00:%x[-2,0]

U10:%x[-2,1]

D. Formation of the file for model training and testing

Set of words from the pre-processed data was selected and added to the training corpus. 95 percentage of the words from the dataset were taken for training and the remaining words were kept for testing. The words in the training file were added with named entity tags depending on the named entity class to which they belong to. BIO notation was followed to perform named entity tagging. First word of a particular named entity was marked as 'B' followed by the particular class of the named entity, all the remaining words were tagged as 'I' followed by the particular named entity class and 'O' denotes out of the named entity or not a named entity.
For example if a person`s name in corpus is 'Ramachandra Karunakara Menon'

The named entity tagging will be as follows:
Ramachandra B-INDIVIDUAL
Karunakara I-INDIVIDUAL
Menon I-INDIVIDUAL.

Two sets of test corpuses were formed each of 5000 words. The test corpus I was formed with the remaining 5000words available with the initial dataset collected. The test corpus II was formed with words from tourism domain itself but those which was extracted from another unrelated dataset. Those words were classified into appropriate named entity class for the process of result analysis and identification of the performance of the system, but the named entity tags were not added in the test files.

E. Training the model

The training file together with the file containing feature templates was used to train and form the model. The machine learning approach used here is Conditional Random field approach which was explained in section III. CRF model generated the model file which includes the learnt model. The next stage of testing and result analysis is discussed in Section V.

V. TESTING AND RESULT ANALYSIS

The test corpus was given as input to the trained model and it performs the process of named entity tagging. The results were analyzed based on Precision, Recall and F-measure. Precision measures the numbers of named entities that are tagged correctly out of the named entities tagged, recall measures the number of named entities appropriately tagged out of all the named entities present in the corpus and F-measure was calculated as the harmonic mean of recall and precision. The results are displayed in table II.

TABLE II: RESULTS

Test Dataset	Precision in %	Recall in %	F-measure in%
Test Dataset I	95.12	62.33	75.31
Test Dataset II	85	50	62.96

The result shows that combining rule-based approaches and machine learning based approach can make the system more efficient. Also, though the test corpus II was extracted from totally unrelated dataset from the same domain still the performance of the model is reasonable.

VI. CONCLUSION AND FUTURE WORK

NER systems for Indian languages being still in its infancy stage, the proposed system can open a lot of scope for further research in this area for Malayalam as well as other highly agglutinative languages. The proposed model performs the process of NER with reasonably good performance. As part of future work more rules based on the morphology of the language can be formed and tested to develop NER systems for Indian languages as efficient as NER systems for English language. More resources like word list, word nets etc. can also be collected and used for development of NER systems.

REFERENCES

1. Gowri Prasad, K.K. Fousiya, M. Anand Kumar, K.P. Soman,"Named Entity Recognition for Malayalam language: A CRF based approach" in the proceedings of 2015 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials(ICSTM), 2015 pages 16-19.
2. Rayner Alfred, Leow Chin Leong, Chin Kim On, and Patricia Anthony, "Malay named entity recognition based on rule-based approach" in International Journal of Machine Learning and Computing, Vol. 4, No. 3, June 2014
3. I. Budi, S. Bressan, "Association Rules Mining for Name Entity Recognition",Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003.
4. Collins, Michael and Y. Singer. "Unsupervised models for named entity classification", In proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.
5. J. Kim, I. Kang, k. Choi, "Unsupervised Named Entity Classification Models and their Ensembles",*Proceedings of the 19th international conference on Computational linguistics*, 2002.
6. Shaodian Zhang and Noemie Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts" *in Elsevier-Journal of Biomedical Informatics* 46, pages 1088-1098 , August 2013
7. D.M. Bikel, S. Miller, R. Schwartz and R, Weischedel, "A high-performance learning name-finder", fifth conference on applied natural language processing, PP 194-201, 1998.
8. G.D Zhou and J.Su (2002) "Named entity recognition using an hmm-based chunk tagger," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 473-480.
9. F. Bechet, A. Nasr and F. Genet, "Tagging Unknown Proper Names Using Decision Trees", In proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000.
10. Y. C. Wu, T.K Fan, Y. S Lee and S. J Yen (2006) ``Extracting Named Entities Using Support Vector Machines", Spring-Verlag, Berlin Heidelberg, 2006.
11. Kishorjit Nongmeikapam, Laishram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Ngariyanbam Mayekleima Chanu and Sivaji Bandyopadhyay. 2011. "CRF based named entity recognition in manipuri: a highly agglutinative language". Proceeddings of 2nd National Conference on Emerging Trends and Applications in Computer Science, March 2011
12. VijayaKrishna R. and Sobha L. "Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields". *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pages 59--66, Hyderabad, India, January 2008.
13. Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar and Pabitra Mitra(2008) "A hybrid approach for named entity recognition in indian languages," Proceedings of the IJCNLP-08 .
14. Gowri Prasad, K.K. Fousiya, "Named Entity Recognition approaches: A study applied to English and Hindi language" in the proceedings of 2015 IEEE International Conference on Circuits, Power and Computing technologies (ICCPCT), 2015 pages 1-4.
15. Jisha P Jayan, Rajeev R.R and Elizabeth Sherly(2012), "A hybrid statistical approach for named entity recognition for malayalam language," International Joint Conference on Natural Language Processing, pages 58-63, Nagoya, Japan, 14-18 October 2013.
16. Bindu.M.S and Sumam Mary Idicula. " Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods*", in the International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3*, pages 185-191, September 2011.

AUTHORS PROFILE

Gowri Prasad is working as Assistant Professor in Department of Information Science and Engineering, New Horizon college of Engineering, Bengaluru. Her area of research is Machine Learning and Artificial Intelligence.

Prakruthi S T is working as Assistant Professor in Department of Information Science and Engineering, New Horizon college of Engineering, Bengaluru. Her area of research is Machine Learning.