Named Entity Recognition for Malayalam language: A CRF based approach

Conference Paper · May 2015

DOI: 10.1109/ICSTM.2015.7225384

CITATIONS
6

READS
557

4 authors, including:

M. Anand Kumar
National Institute of Technology Karnataka
115 PUBLICATIONS **656** CITATIONS
SEE PROFILE

Soman Kp
Amrita Vishwa Vidyapeetham
656 PUBLICATIONS **4,217** CITATIONS
SEE PROFILE

2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.16-19.

Named Entity Recognition for Malayalam Language: A CRF based Approach

Gowri Prasad[1], K.K. Fousiya[2], Dr. M. Anand Kumar[3] and **Dr. K.P. Soman[4]**

[1,2]Computer Science And Engineering, Jyothi Engineering College, Thrissur, India
[3,4]Center for Excellence in Computational Engineering and Networking, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore.
Email: [1]gowri.0588@gmail.com, [2]fousiyakk@jecc.ac.in

***Abstract*— Named Entity Recognition is an important application area of Natural Language Processing. It is the process of identifying the designators which are present in a sentence called as named entities. Named Entity Recognition can be performed using rule based approaches, machine learning based approaches and hybrid approaches. This paper proposes a method for Named Entity Recognition of Malayalam language using one of the supervised machine learning approach called Conditional Random field approach.**

***Keywords*— Named Entity Recognition, Machine learning, Supervised Machine Learning, Conditional Random Field Approach.**

I. INTRODUCTION

Named Entity Recognition (NER) performs the process of identification of the designators present in a sentence called the named entities like person names, places, organisation, date, numerical expressions etc. NER can be performed mainly using two steps viz. identification and classification. In short NER systems identify the named entities and classify them into its appropriate name class. NER is used as part of information extraction systems, question answering systems, summarisation systems, search engines etc and it is also used for simplifying tasks like machine translation, aggregation of documents, automatic indexing of books etc.

NER is a very difficult task due to the ambiguities present in natural language. A word can be member of more than one name class based on the context in which they are used. An example for this is the word 'Yamuna' which can be a person name or the name of river. Another example for ambiguity is the word 'Adoor Gopalakrishnan'. Here 'Adoor Gopalakrishnan is a person name while 'Adoor' is also a place name. There are efficient NER systems in languages like English which gives high f-measure values. Many languages especially Malayalam and other Indian languages lack efficient NER systems. The challenges faced in performing NER for Malayalam is due to many reasons. One of the main reason is the lack of resources like annotated corpus, lists, name dictionaries etc. Malayalam language is highly agglutinative in nature and also lacks capitalisation information. It does not have subject verb object agreement and so is free word order in nature. It is also morphologically rich language.

NER can be performed using different approaches like rule based approaches, machine learning based approaches or other hybrid approaches that combines rule based and machine learning based strategies. Rule based approaches are highly language dependent as it depends on certain rules formed by linguists to identify named entities. It also uses lists and such other resources which are specific to a language. Machine learning based approaches can be supervised, semi-supervised or unsupervised. In semi-supervised and unsupervised approaches much human intervention is not required as it applies the strategies of pattern recognition, automated signature generation based on context words etc. But in supervised machine learning based approach annotated corpus is used for training before NER is performed.

This paper proposes a supervised machine learning based approach for NER using CRF model. Here section II describes the related works in the same area and section III explains briefly about CRF model. This followed by section IV which describes in detail the proposed system for NER in Malayalam. Section V gives the results and the result analysis followed by conclusion in section VI.

II. RELATED WORKS

Rule based approach for NER has been tried in many different languages. This approach is proposed in [1] and [2]. Unsupervised approach for NER using an untagged corpus is described in [3] and [4]. Unsupervised approach for NER has been applied specifically to biomedical domain

in [5]. Supervised approach for NER is used for many languages especially those which are rich in resources. Supervised machine learning based NER systems using hidden markov model is proposed in [6], [7] and, in [8] NER using support vector machine is explained. In [9] the authors suggest that NER can also be performed using decision trees.

NER systems have been developed for many Indian languages. CRF based NER was experimented in [10] and [11] for Manipuri and Tamil languages respectively. A hybrid approach for NER combining rule based and machine learning based approach has been suggested in [12]. NER systems for Malayalam language was proposed in [13] and [14]. Different approaches for supervised machine learning based NER is suggested in [15][16][17]

III. CONDITIONAL RANDOM FIELDS

In the proposed system for NER we have used conditional random field based model [15]. Generative models assign joint probability for observation and sequences of labels. Hidden Markov Model is a typical example of a generative model. Discriminative or conditional models assign a probability for possible sequence of label given an observation sequence. Labeling depends on the current, past and future observation. CRF based model is an example of such a model. CRF`s build a single exponential model to determine the joint probability of sequences of labels given the entire observation sequence. It is a model which does not have per state normalization of Transition Probabilities and is globally conditioned on the observation. Main advantage of CRF based model is that it solves the problem of label bias. We have used CRF++ 0.58 package[1] which is available as open source for this work.

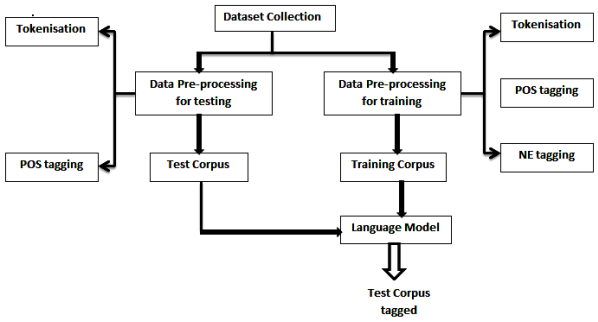

Fig. 1. The System for NER in Malayalam

IV. NER FOR MALAYALAM

The proposed system does the process of NER by going through different stages. The different stages are shown in Fig 1. First the corpus or dataset used for preparation of training and testing files is collected. Features which are used for training the model is also decided. Then the training and testing files are prepared. The training file is prepared through the process of tokenization, part of speech tagging and NE tagging. The other features are also added to the training file. The test file should also be of the same format as the training file except that NE tags will not be present in the test file. The feature template file is also created. Then the training is done to create the model and the test file is given as input to the model for NE tagging. Finally we get tagged corpus.

A. Dataset Collection and Data Pre-Processing

Our corpus consists of almost 50000 words in Malayalam from tourism domain. It was collected from news articles, data from magazines of tourism domain, websites etc. Some part of dataset was also collected from corpus for NER Fire. The collected dataset was tokenised to split the sentences into individual words. Part of Speech tagging was done to this tokenised text for tagging each word with its part of speech category. Chunk tags were also added for each word. Test corpus was also prepared by performing the process of tokenisation, part of speech tagging and chunk tagging.

Table I. Tagset

Tag Assigned	*Description*
B-INDIVIDUAL	Beginning word and succeeding words of person name
I-INDIVIDUAL	
B-LOCATION	Beginning word and succeeding words of place name
I-LOCATION	
B-DATE	Beginning word and succeeding words of date
I-DATE	
B-ORGANISATION	Beginning word and succeeding words of organisation names
I-ORGANISATION	
B-COUNT	Beginning word and succeeding words of counts
I-COUNT	
B-QUANTITY	Beginning word and succeeding words that describe quantities
I-QUANTITY	
B-DISTANCE	Beginning word and succeeding words that describe distances
I-DISTANCE	
B-PERIOD	Beginning word and succeeding words that describe periods
I-PERIOD	
B-WATERBODIES	Beginning word and succeeding words of names of rivers, lakes etc.
I-WATERBODIES	
B-FRUIT	Beginning word and succeeding words of fruits
I-FRUIT	

NE tagging was also done for the words in the training corpus. Those words which can be classified as a named

entity was given the appropriate tag. NE tagging was done using BIO notation as shown in table I i.e. first word of a NE will be tagged as 'B-abc' and the following words of the NE will be tagged as 'I-abc', where 'abc' is the NE tag assigned for that particular name class. 'o' is the tag given for those words which cannot be classified as a named entity. Thus our training file consists of the word, the part of speech tag, the chunk tag and the last column as NE tag, each separated by a white space in each row. Test corpus is also of the same format as training corpus except that it does not have the NE tag as the last column.

B. Feature Template File

Features are used to train the model. Different feature like Part of speech features, n-gram features, binary features, morphological features etc. can be used for training the model. In our proposed system the features we used are the basic word and Part of speech features. They are the part of speech of the previous two words, current word and succeeding two words, chunk tags of previous two words, current word and succeeding two words, previous two words and succeeding two words as context words, current word and combinations of each of the features specified etc. An extract from feature template file which contains only unigram features is shown in Fig 2.

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]
```

Fig. 2. An Extract From Feature Template File

C. Training and Testing

Training is done using the training file and feature template file. CRF model learns using those files and generates the model file which contains the learnt model. During testing we give this model file as input together with the test file containing the test corpus which contains one less than the number of fields in training files i.e. the NE tags. After testing we get output file in the same format as training file with the corresponding named entity tag in the last column. The named entities are tagged as trained using BIO notation as explained in section IV A.

First training was done using a training file having 25000 words and tested on test corpus of 2000 words. Then training was repeated using the same training file with all the words in training corpus and tested on same test corpus.

V. Results and Result Analysis

The result obtained from the system for NER is analysed in terms of three parameters which are precision, recall and F-measure. The results

- *Precision*— Percentage of selected NEs that are identified correctly and tagged properly.
- *Recall*— Percentage of number of NEs correctly selected out of total number of NEs.
- *F-Score*— A measure that assesses the P/R trade-off which is the harmonic mean of precision and recall.

The results are given in table II. Testing has been done on the same test corpus during both rounds of training. We can see that the training done using the training corpus with more number of words are more efficient and gives higher F-measure values.

Table II. Results of NER Performed

Training Corpus (word count)	*No of NE`s in test corpus*	*Precision (%)*	*Recall (%)*	*F-measure (%)*
25000	566	83.5	41.7	55.6
30000	549	90.9	62.9	74.35

VI. Conclusion and Future Work

Malayalam being one of the oldest languages lacks an efficient NER system and so the growth in the field of information extraction for Malayalam language is very slow. The proposed system does the process of NER with reasonable efficiency in terms of F-measure values. We can also conclude that for supervised machine learning based approaches the efficiency is highly dependant on the accuracy and the number of words in the training corpus. F-measure values are more when training corpus consists of more number of words.

As part of future work we intend to collect more dataset. We are also planning to add more features for training

which will increase the f-measure values further. We also intend to test using other supervised machine learning tools.

REFERENCES

[1] 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

[2] I. Budi, S. Bressan, "Association Rules Mining for Name Entity Recognition",Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003.

[3] 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.

[4] J. Kim, I. Kang, k. Choi, "Unsupervised Named Entity Classification Models and their Ensembles",***Proceedings of the 19th international conference on Computational linguistics***, 2002.

[5] 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

[6] 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.

[7] 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.

[8] 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.

[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] 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

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] John Lafferty, Andrew McCallum, Fernando CN Pereira, "Conditional Random Field- A probabilistic model for segmenting and labelling sequence data***" in proceedings of the 18 th International Conference on Machine Learning 2001***, pages 282-289

[16] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. "Exploiting diverse knowledge sources via maximum entropy in named entity recognition" 1998

[17] Marti A. Hearst, "Support Vector Machines", ***IEEE Intelligent Systems,*** pages 18-21, July/August 1998