A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language

Jisha P Jayan
jisha.jayan@iiitmk.ac.in

Rajeev R R
rajeev@iiitmk.ac.in

Elizabeth Sherly
sherly@iiitmk.ac.in

Abstract

Named-Entity Recognition (NER) plays a significant role in classifying or locating atomic elements in text into predefined categories such as the name of persons, organizations, locations, expression of times, quantities, monetary values, temporal expressions and percentages. Several Statistical methods with supervised and unsupervised learning have applied English and some other Indian languages successfully. Malayalam has a distinct feature in nouns having no subject-verb agreement, which is of free order, makes the NER identification a complex process. In this paper, a hybrid approach combining rule based machine learning with statistical approach is proposed and implemented, which shows 73.42% accuracy.

1 Introduction

Named-Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the name of persons, organizations, locations, expression of times, quantities, monetary values, temporal expressions, percentages, etc. There are different supervised and unsupervised learning approaches for NER using statistical methods like HMM, Decision Forest, Maximum Entropy, SVM, Conditional Random fields etc. The term Named Entity was introduced in the sixth Message Understanding Conference (MUC-6). In fact, the MUC conferences were the events that have contributed in a decisive way to the research of this area. It has provided the benchmark for named entity systems that performed a variety of information extraction tasks (Mansouri et al., 2008).

The named entities are generally nouns. NER although a seemingly simple task, but a difficult task to find, and once found, difficult to classify. For example, locations and person names can be the same, and follow similar formatting. NEs are typically not registered in general-purpose lexical resources while generic terms are expressed. NEs are subject to permanent changes and show syntactic behaviour which is specific to them. NEs, generic terms and its various forms are used interchangeably and form chains of co-referring items.

Malayalam belongs to the Dravidian family of languages and is one of the four major languages of this family with a rich literary tradition, inflectionally adding of suffixes with the root or the stem word forms rich in morphology. The language must be certainly being older, but linguistic research is yet to be discovering unmistakable evidence to prove its antiquity. NER tasks are still difficult and in infancy in many Indian languages, and is more in Malayalam.

2 Related Works

In recent years, automatic named entity recognition and extraction systems have become one of the popular research areas that a considerable number of studies have been addressed on developing these systems. They can be categorized into three classes namely, Rule based NER, Machine Learning based NER and Hybrid based NER (Wu et al., 2006). Hand-made or Rule-based focuses on extracting names using human-made rules set.

Generally the system consist of set of patterns using grammatical, syntactic and orthographic features in combination with dictionaries (Budi et al., 2003). These approaches are relying on manually coded rules and manually compiled corpora. These kinds of models have better results for restricted domains, are capable of detecting complex entities that learning models have difficulty with. However, the rule-based NE systems lack the ability of portability and robustness, and furthermore the high cost of the rule maintains increases even though the data is slightly changed. These type of approaches are often domain and language specific and do not

International Joint Conference on Natural Language Processing, pages 58–63,
Nagoya, Japan, 14-18 October 2013.

necessarily adapt well to new domains and languages.

Generally the system consist of set of patterns using grammatical , syntactic and orthographic features in combination with dictionaries (Budi et al., 2003). These approaches are relying on manually coded rules and manually compiled corpora. These kinds of models have better results for restricted domains, are capable of detecting complex entities that learning models have difficulty with. However, the rule-based NE systems lack the ability of portability and robustness, and furthermore the high cost of the rule maintain increases even though the data is slightly changed. These type of approaches are often domain and language specific and do not necessarily adapt well to new domains and languages.

There are two types of machine learning models that are used for NER called Supervised and Unsupervised machine learning model. Supervised learning involves using a program that can learn to classify a given set of labeled examples that are made up of the same number of features. Each example is thus represented with respect to the different feature spaces. The learning process is called supervised, because the people who marked up the training examples are teaching the program the right distinctions.

In recent years several statistical methods based on supervised learning method were proposed. Bikel et. al. proposed a learning name-finder based on hidden Markov model (Bikel et al. , 1998) called Nymbel, while Borthwick et. al. investigates exploiting diverse knowledge sources via maximum entropy in named entity recognition (Borthwick et al. , 1998).

A tagging of unknown proper names system with Decision Tree model was proposed by Bechet et. al. (2000), while Wu et. al. (2006) presented a named entity recognition system based on support vector machines.

Unsupervised learning method is another type of machine learning model, where an unsupervised model learns without any feedback. In unsupervised learning, the goal of the program is to build representations from data. These representations can then be used for data compression, classifying, decision making, and other purposes. Unsupervised learning is not a very popular approach for NER and the systems that do use unsupervised learning are usually not completely unsupervised. In these types of approach, Collins et. al.(1999) discusses an unsupervised model for named entity classification by use of unlabeled examples of data.

Koim et. al. (2002) proposed an unsupervised named entity classification models and their ensembles that uses a small-scale named entity dictionary and an unlabeled corpus for classifying named entities. Unlike the rule- based method, these types of approaches can be easily port to different domain or languages.

VijayaKrishna et al. (2008) also experimented with Conditional Random Field (CRF) models for a domain focused Tamil Named Entity Recognizer for tourism domain. Their observation resulted that Conditional Random Fields is well suited for Named Entity recognition for Indian languages, but it is tested only for the noun phrases.

Sujan Kumar Saha et al. of IIT, Kharagpur used a hybrid approach for their NER task in Indian Languages. The hybrid techniques include Maximum Entropy model (MaxEnt), language specific rules and gazetteers. For their work they have considered 5 Indian languages – Hindi, Bengali, Oriya, Telugu and Urdu.

Kishorjit Nongmeikapam et al. (2011), has explored the NER task for Manipuri in their work - CRF Based Name Entity Recognition (NER) in Manipuri: a highly agglutinative Indian Language using Conditional Random Field (CRF).

In Hybrid NER system, the approach is to combine rule- based and machine learning-based methods, and make new methods using strongest points from each method. In this family of approaches Mikheev et. al. proposed a Hybrid document centered system, called LTG system (Mikheev et al. , 1998) Sirihari et. al.(2000) introduced a hybrid system by combination of HMM, MaxEnt, and handcrafted grammatical rules.

Statistical methods work by using a probabilistic model containing features of the data which are similar to the rule-based approaches. The features of the data, which could be understood as rules set for the probabilistic model, are produced by learning the resulting corpora with correctly marked named entities. The probabilistic model then uses the features to calculate and identify the most probable named entities. As such, if the annotated features of the data are truly reliable, the model would have a high probability in finding almost all the named entities within a text.

3 Statistical Apporach

The statistical (Brants, 2000) methods are mainly based on the probability measures including the unigram, bigram, trigram and n-grams. TnT- Trigrams n Tags is a very efficient statistical part of speech tagger that can be trained on any language with any tagset. The parameter generation component trains on tagged corpora. The system uses several techniques for smoothing and handling of unknown words. TnT can be used for any language, adapting the tagger to a new language, new domain or new tagset very easy.

The tagger is implemented using Viterbi algorithm for second order Markov models. Spanish TnT is a statistical approach, based on a Hidden Markov Model that uses the Viterbi algorithm with beam search for fast processing.

The Viterbi algorithm is used to compute the most likely tag sequence in O(W x T2) time where T is the number of possible part-of-speech tags and W is the number of words in the sentence. It performs the maximum likelihood probability calculation using the parameters from lexicon file and n-gram file. The algorithm sweeps through all the tag possibilities for each word computing the best sequence leading to each possibility. The key that makes this algorithm efficient is that the usage of best sequences leading to the previous word because of the Markov assumption.

TnT is trained with different smoothing methods and suffix analysis. The parameter generation component trains on tagged corpora. The system uses several techniques for smoothing and handling of unknown words. Linear interpolation is the main paradigm used for smoothing and the weights are determined by deleted interpolation. To handle the unknown words, suffix trie and successive abstraction are used.

TnT's greatest advantage is its speed, important both for fast tuning cycle when dealing with large corpora. The strong side of TnT is its suffix guessing algorithm that is triggered by unseen words. From the training set TnT builds a trie from the endings of words appearing less than n times in the corpus, memorizes the tag distribution for each matrix. A clear advantage of this approach is the probabilistic weighting of each label, however, under default settings the algorithm proposes a lot more possible tags than a morphological analyzer would.

4 Proposed Work

Malayalam language treats the named entities as Nouns and so they are Noun Phrases. All Noun Phrases are not named entities and can have morphological inflections as Malayalam is morphologically rich. This makes a single named entity to appear as different words in different context. Malayalam lacks capitalization information for named entities and one named entity can appear with different meaning in another context. For example, consider the word 'Kavitha' is a common noun with the meaning name of a person and 'poem' and also a Proper Noun. The free word order of the language is also posing problems as NEs can appear in subject and object positions. The language construct has no Subject-Verb agreement and there exists a free word order so that named entities can appear in any position. Therefore, Malayalam requires properly tailored method for identification of NER and we propose a supervised machine learning method using TnT based on a Hidden Markov Model and Viterbi algorithm.

5 Implementation

The major steps involved in NER are Corpus selection, POS Tagging, NER Tagging, training the corpus using the TnT to create lexicon and the ngram files. Based on these language models generated, the raw corpus with POS annotation are tagged. Rules are used in some cases where there occurs the inner and outer tags. The architecture used for recognizing the named entities in Malayalam is shown in Figure1.

5.1 Tagging

The Named entity hierarchy is divided into three major classes; Entity Name, Time and Numerical expressions. The Name hierarchy has eleven attributes. Numeral Expression and time have four and three attributes respectively. Named Entities are tagged using the tagset developed for Indian Language Machine Translation and CLIA projects of DEITY, Government of India. This tagset is hierarchical in nature and the first level tags consist of ENAMEX, TIMEX and NUMEX. The first level tags of ENAMEX consists of 11 tags with 46 subtags and 20 tags under the subtags . NUMEX has 4 subtags whereas TIMEX has 7 subtags.

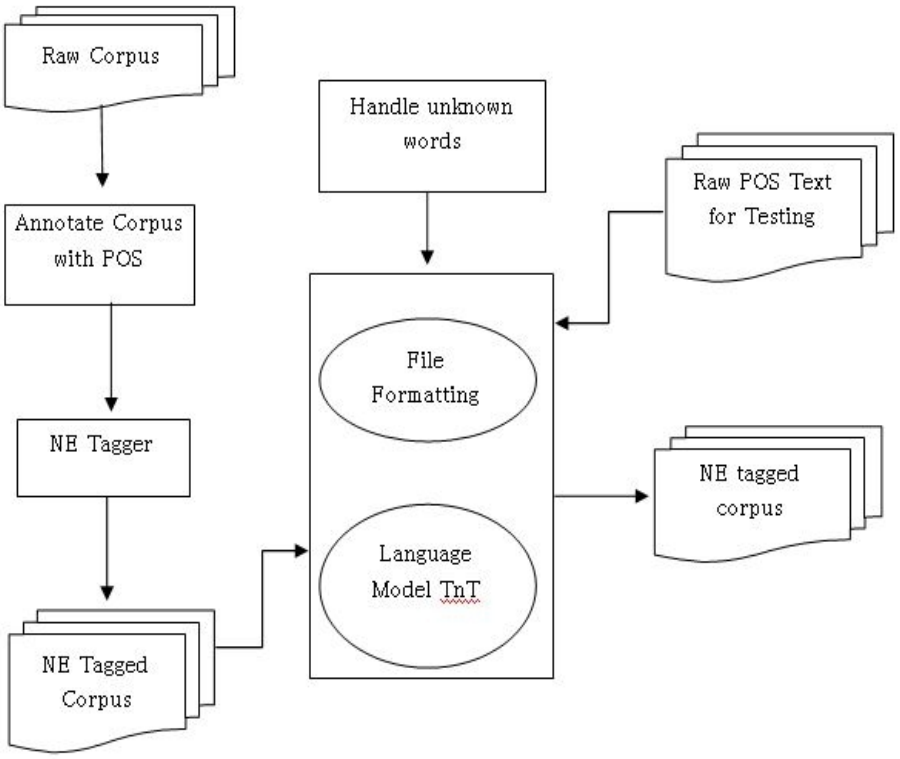

Figure 1: NER Architecture for unknown words

Examples

<ENAMEX TYPE="LOCATION" SUBTYPE_1="PLACE" SUBTYPE_2="STATE">
1.1 കേരളത്തിന്റെ NNP
</ENAMEX>
<ENAMEX TYPE="LOCATION" SUBTYPE_1="WATERBODIES">
<ENAMEX TYPE="LOCATION" SUBTYPE_1="PLACE">
2.1 ബേക്കൽ NNP
</ENAMEX>
2. ഫാള്ാ: NN
</ENAMEX>

There are several occasions where embedded tags are used. For example:
ഈന്ററ്റക്ക ഈക്കുത്തീച്ചുറ്റ് ഔബ്
ഈക്കുബച്ചയേസക്ക ഠെഖ്പേിഴഝീ ഇക്കുഡ്
യിപേഝ്യെമ്മ് ഖേറഴ

(Indian Institute of Information Technology and Management-Kerala), where "ഈന്ററ്റക്ക (Indian) " takes the tag GPE and "ഖേറഴ (Kerala)" takes the tag State while other tokens take no NER tags, but as a whole this refers to an Institute with its Tag Institute under Facility. In these cases, there occurs the need for writing the rules to identify the same as an Institute. The hybrid approach is more useful in such cases.

6 Experiments and Results

Under the same domain, a comparison on two supervised taggers namely TnT and SVM was conducted. In our experiment, for known words, SVM shows better performance but for unknown words TnT outperformed. However, for embedded tags, it is required to generate rules that combining with TnT shows better result. So our proposed hybrid supervised machine learning approach with the combination of TnT and Rule based is a good strategy for NER especially for embedded tags.

The corpus was tagged using the NER tagset for Malayalam. The TnT was learned using the tagged corpus. When learned, the dictionary file was created for the corpus. Once learning process is done, then the input text file was given to the tool and tagging was performed. The system gives an accuracy of 73.42% .

Size of training corpus (in tokens)	Size of test corpus (in tokens)	Automated accuracy obtained	Precision (%)	Recall (%)	F-Measure
100	150	57.59%	37.5	26.09	30.77
200	150	56.96%	56.25	39.13	46.15
500	150	60.76%	58.33	30.43	39.10
2000	150	68.99%	87.5	30.43	45.16
5000	150	73.42%	100	30.43	46.66
10000	150	73.42%	100	43.48	60.61

Table 1: **Result of NE tagging using TnT**

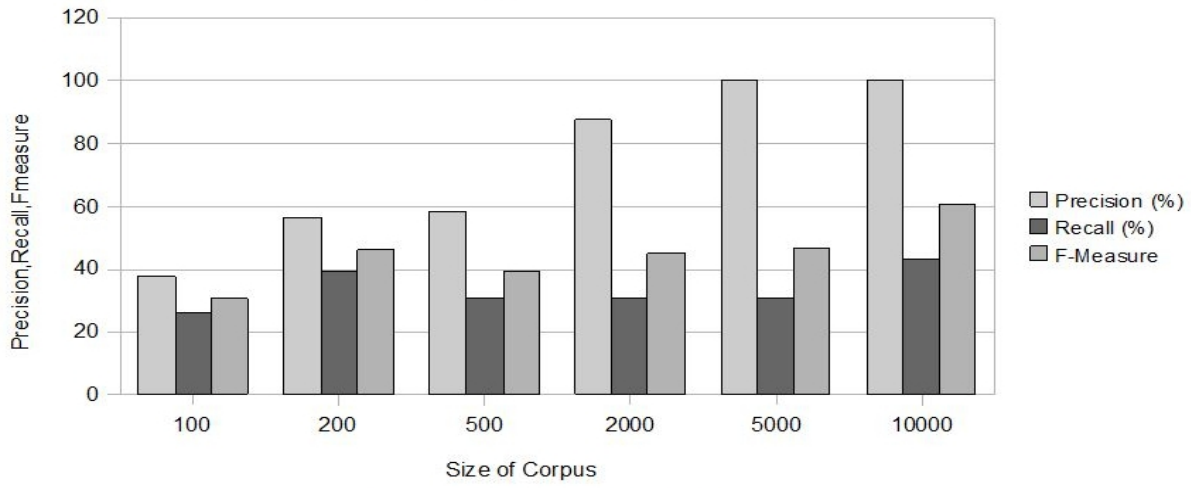

Figure 2 : Precision, Recall and F-Measure analysis

The accuracy can be increased by increasing the amount of training data. The detailed observations are given in the Table 1. The performance of NER system in Malayalam is computed based on the parameters-Precision, Recall and F-Measure. Recall is defined as the number of correct tags in the document marked up by our proposed NER system over the total number of annotated-tags present in the document. The main purpose of recall is to measure how well our system can perform the recognition of entity names. Precision is defined as the number of correct tags in the file marked up by our system over the total number of tags being marked up.

Precision=Correct NEs / Total NEs identified by the System

Recall = Correct NES/ Gold standard NEs in the System

7 Conclusion

Many natural language processing applications require finding Named Entities in textual documents. Named Entity Recognition plays a significant role in various language processing applications such as Question Answering and Summarization Systems, Information Retrieval, Machine Translation, Video Annotation, Semantic Web Search and Bioinformatics. Considering the various issues like classifying ambiguous strings correctly, detecting the boundaries of an NE correctly, categorizing NERs, and availability of Unicode data, the proposed hybrid model achieves 73.42% accuracy. The domains considered for tagging were health and tourism. Accuracy can be further increased by increasing the number of words in the training corpus. The work shows that a hybrid statistical approach, combining TnT and rule based suit better for highly morphologically and inflectionally rich languages like Malayalam.

References

Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat,IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008

Y.C. Wu, T.K. Fan, Y.S. Lee, S.J Yen, ``Extracting Named Entities Using Support Vector Machines", Spring-Verlag,Berlin Heidelberg, 2006.

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

D.M. Bikel, S. Miller, R. Schwartz, R, Weischedel, "a High-Performance Learning Name-finder", fifth conference on applied natural language processing, PP 194-201, 1998.

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

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.

Y.C. Wu, T.K. Fan, Y.S. Lee, S.J Yen, ``Extracting Named Entities Using Support Vector Machines", Spring-Verlag, Berlin Heidelberg, 2006.

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.

Sujan Kumar Saha,Sanjay Chatterji,Sandipan Dandapat,Sudeshna Sarkar,Pabitra Mitra 2008.A Hybrid Approach for Named Entity Recognition in Indian Languages, Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 17--24, Hyderabad, India, January , Asian Federation of Natural Language Processing

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

A. Mikheev, C. Grover, M. Moens, "Description OF THE LTG SYSTEM FOR MUC-7", In Proceedings of the seventh Message Understanding Conference (MUC-7), 1998.

R. Sirhari, C. Niu, W. Li, "A Hybrid Approach for Named Entity and Sub-Type Tagging" Proceedings of the sixth conference on Applied natural language processing ,Acm Pp. 247 - 254 , 2000.

T. Brants. TnT --- A Statistical Part of-Speech Tagger. In Proceedings of the 6th Applied NLP Conference (ANLP-2000), pages 224--231, 2000.

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2009.Arabic Named Entity Recognition: A Feature-Driven Study. IEEE Transactions on Audio, Speech, and Language Processing, VOL. 17, NO. 5, July 2009

Kishorjit Nongmeikapam, Laishram Newton Singh, Tontang Shangkhunem, Bishworjit Salam, Ngariyanbam Mayekleima Chanu, 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

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.