Tamil NER – Coping with Real Time Challenges

Malarkodi, C S., Pattabhi, RK Rao and Sobha, Lalitha Devi
AU-KBC RESEARCH CENTRE, MIT Campus of Anna University, Chrompet, Chennai, India
csmalarkodi@au-kbc.org,pattabhi@au-kbc.org,sobha@au-kbc.org

ABSTRACT

This paper describes various challenges encountered while developing an automatic Named Entity Recognition (NER) using Conditional Random Fields (CRFs) for Tamil. We also discuss how we have overcome some of these challenges. Though most of the challenges in NER discussed here are common to many Indian languages, in this work the focus is on Tamil, a South Indian language belonging to Dravidian language family. The corpus used in this work is the web data. The web data consisted of news paper articles, articles on blog sites and other online web portals.

KEYWORDS: Named Entity Recognition, NER, Challenges, Features, post-processing

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38,
COLING 2012, Mumbai, December 2012.

1 Introduction

This Named Entity Recognition (NER) refers to automatic identification of named entities in a given text document. NER refers to the recognition and classification of proper nouns in the given document. NER can be viewed as labeling task. Given a text document, named entities such as Person names, Organization names, Location names, Product names are identified and tagged. Identification of named entities is important in several higher language technology systems such as information extraction systems, machine translation systems, and cross-lingual information access systems.

Named Entity Recognition was one of the tasks defined in MUC 6. Several techniques have been used for Named Entity tagging. A survey on Named Entity Recognition was done by David Nadeau (2007). The techniques used include rule based technique by Krupka (1998), using maximum entropy by Borthwick (1998), using Hidden Markov Model by Bikel (1997) and hybrid approaches such as rule based tagging for certain entities such as date, time, percentage and maximum entropy based approach for entities like location and organization (Rohini et al., 2000) There was also a bootstrapping approach using concept based seeds (Niu et al., 2003) and using maximum entropy markov model (Finkel et al., 2004). Alegria et al, (2006), have developed NER for Basque, where NER was handled as classification task. In their study, they have used several classification techniques based on linguistic information and machine learning algorithms. They observe that different feature sets having linguistic information give better performance.

Lafferty (2001) came up with Conditional Random Fields (CRFs), a probabilistic model for segmenting and labeling sequence data and showed it to be successful with POS tagging experiment. Sha and Pereira (2003) used CRFs for shallow parsing tasks such as noun phrase chunking. McCallum and Li (2003) did named entity tagging using CRFs, feature induction and web enhanced lexicons. CRFs based Named Entity tagging was done for Chinese by Wenliang Chen (2006). CRFs are widely used in biological and medical domain named entity tagging such as work by Settles (2004) in biomedical named entity recognition task and Klinger's (2007) named entity tagging using a combination of CRFs. The Stanford NER software (Finkel et al., 2005), uses linear chain CRFs in their NER engine. Here they identify three classes of NERs viz., Person, Organization and Location. Here they have used distributional similarity features in their engine, but this utilizes large amount of system memory.

In Indian languages many techniques have been used by different researchers. Named Entity recognition for Hindi, Bengali, Oriya, Telugu and Urdu (some of the major Indian languages) were addressed as a shared task in the NERSSEAL workshop of IJCNLP. The tagset used here consisted of 12 tags. In this shared task different research groups had participated. The groups had used techniques such as SVM, CRFs, MEMM, and rule based approaches. Vijayakrishna & Sobha (2008) worked on Domain focused Tamil Named Entity Recognizer for Tourism domain using CRF. It handles nested tagging of named entities with a hierarchical tag set containing 106 tags. They considered root of words, POS, combined word and POS, Dictionary of named entities as features to build the system. Pandian et al (2007) have built a Tamil NER system using contextual cues and E-M algorithm.

The NER system (Gali et al., 2008) build for NERSSEAL-2008 shared task which combines the machine learning techniques with language specific heuristics. The system has been tested on five languages such as Telugu, Hindi, Bengali, Urdu and Oriya using CRF followed by post processing which involves some heuristics. Ekbal & Bandyopadhyay (2009) had developed Named Entity Recognition (NER) systems for two leading Indian languages, namely Bengali and Hindi using the Conditional Random Field (CRF) framework. The system makes use of different types of contextual information along with a variety of features that are helpful in predicting the different named entity (NE) classes. They have considered only the tags that denote person names, location names, organization names, number expressions, time expressions and measurement expressions.

Shalini & Bhattarcharya (2010) developed an approach of harnessing the global characteristics of the corpus for Hindi Named Entity Identification using information measures, distributional similarity, lexicon, term co-occurrence and language cues. They described that combining the global characteristics with the local contexts improves the accuracy. The work proposed by Kumar, et al. (2011) to identify the NEs present in under-resourced Indian languages (Hindi and Marathi) using the NEs present in English, which is a high resourced language. The identified NEs are then utilized for the formation of multilingual document clusters using the Bisecting k-means clustering algorithm.

In this paper we discuss various challenges we faced while developing a NER system. The paper is further organized as follows. Section 2, describes the base NER system using CRFs. In section 3, the various challenges are discussed. Section 4 describes experiments and results. Section 5 concludes the paper.

2 NER system using CRFs

The basic goal of our work is to develop a practical NER system that can be used reliably on the web data. For an automatic system to be used in real time it should be robust. In HMM as the current label depends on the previous one it suffers from dependency problem. It also needs a large training corpus otherwise it has a data sparsity problem. MEMM have labeling bias problem because the probability transition from any given state must sum to one, and hence it has bias towards states with outgoing transitions. Here we have chosen CRFs because of its advantages over HMM and MEMM. CRFs are one best suited technique for sequence labeling task.

2.1 NER Tagset

We have used a hierarchical tagset consisting 106 tags. This hierarchical tagset which consists of mainly three classes Entity Names (ENAMEX), Numerical and Time expressions (NUMEX, TIMEX). In implementation part we had taken 22 second level tags from the hierarchical tagset. Entities may be referenced in a text by their name, indicated by a common noun or noun phrase or represented by a pronoun. Person, organization, Location, Facilities, Cuisines, Locomotives, Artifact, Entertainment, Organisms, Plants and Diseases are the eleven types of Named entities. Numerical expressions are categorized as Distance, Money, Quantity and Count. Time, Year, Month, Date, Day, Period and Special day are considered as Time expressions. This is the standard tagset used widely in Indian Language research community in the national projects such as CLIA and IL-IL MT. Figure 1, shows the partial tagset that has been used in this work.

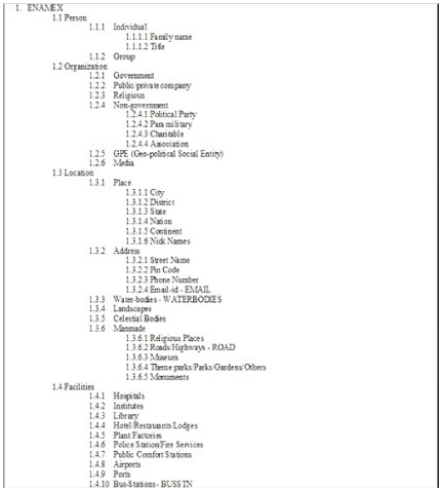

FIGURE 1 – Sample of the NER tagset

2.2 Our NER engine

Capitalization, suffix, prefix, dictionary patterns, first word, frequent word, digit are the widely used features for NE identification. These features vary for each language. Dictionary patterns and gazetteer list varies for every languages and it contains only the frequent named entities. The main goal of our engine is to identify NEs with only basic features such as word, pos and chunk. Current word combined with POS of preceding two words and current word combined with succeeding two words is taken as features. The features used are linguistically inspired. Features used in our system are as follows.

- Word – individual words
- Part of speech tag – pos tag of words
- Combination of word and POS
- Chunk – noun phrase, verb phrase
- Combination of word, POS and chunk

The criteria for the feature selection are explained as follows: POS tag plays a vital role in the task of NE identification and it denotes whether the entities are proper or common nouns or cardinal numbers. The part of speech information helps to identify the relationship between the current, preceding and succeeding words. Phrase chunking is used to find the noun phrases where the named entities have occurred. Combination of features such as word, POS and chunk in the window of five boost the NER engine to learn the context of named entity.

3 Challenges in NER

In the last decade we find that more research interest and work has been done in the area of NER in Indian languages by the research community, yet challenges exist. Indian languages belong to several language families, the major ones being the Indo-European languages, Indo-Aryan and the Dravidian languages. Tamil belongs to Dravidian Language family. The challenges in NER arise due to several factors. In this section we describe some of the main challenges faced by us while developing the system.

3.1 Agglutination

All Dravidian languages including Tamil have agglutinative nature. Case Markers attach as postpositions to proper or common nouns to form a single word.

1. Ta: Koyilil vituwi ullathu
   En: Temple -NN+PSP hostel -NN be-VM
   (there is hostel in the temple)

   In the above example, the case marker “il” suffixed to the common noun “koyil”.

2. Ta: Kerala mannar marthandavarmanukku
   En: Kerala-NN king- NN marthandavarman- NNP+PSP
   therivikkappattathu
   informed-VM
   (informed to the king marthandavarman of kerela)

Where, the case marker “kku” is attached to the proper noun ‘marthandavarman’.

As the case markers suffixed to the noun increases, the number of NEs in the training corpus also increases. So it is difficult for the machine to learn distinct NE patterns.

3.2 Ambiguity

In the corpus we find lot of ambiguity between common and proper nouns. For example the common names such as roja (rose) in Tamil, thamarai (lotus), malar (flower) in dictionaries can also be the names of person and in such cases it should be considered as NE. From the example sentences given below, the problem of ambiguity can be understood easily.

3. Ta: ‘Thiya sakthi thaiva sakthiyai vella mudiyathu’
   En: Evil power god power+acc conquer cannot
   (Evil power cannot overcome the power of god)

Ta:‘Sakthi thanathu tholigalutan koyilukku cenral’

En: Sakthi her with friends temple go+past

(Sakthi went to temple with her friends)

In this example ‘Sakthi” is the ambiguous word having two meanings, one as person name and other as power, or energy.

4. Ta: ‘velli suriyak kudumbathil irandavathaka amainthulla oru kolakum’

   En: Venus solar family second place be-form one planet

   (In the solar system Venus is the second planet from the sun)

   Ta: ‘velli vilai thidirena kuraiththathu’

   En: Silver rate suddenly reduced

   (silver rate suddenly declined)

   Here “velli” has two meanings one name of planet and other as “silver”.

5. Ta: ‘anbu sakalathaiyum thankum’

   En: love all bear

   (Love bears all things)

   Ta: ‘anbu pattap patipai mudithan’

   En: Anbu graduation study complete+past

   (Anbu completed graduation)

   Here “anbu” has two meanings “love” and as name of person.

   In example 4, the word “velli” is an interesting example, where both types are NEs, but belonging to different categories. The first one belongs to category “celestial” and second belongs to “artifact”. Similarly words such as “thamarai” belong to different NE categories “flower” and “person name”.

3.3 Nested Entities

In some cases if the proper noun is considered separately, it belongs to one NE type but if we consider the same NE as a nested entity it belongs to another NE type. For instance,

6. Ta: Kanchi sankaracharyar

   En: kanchi sankaracharya

"kanchi" refers to a location individually, in the case of nested entity it refers to the person "sankaracharyar" who lives in "kanchi".

Ta: andal sannathi

En: andal temple

In general we observe that longer length entities are formed with the combination of two or more individual named entities.

7. Ta: indira ganthi palkalaikallagam - Organization

   En: Indira Gandhi University

   Ta: indira ganthi - Person name

   En: Indira Gandhi

8. Ta: krishna jeyanthi – Special day

   En: Krishna jayanthi (birthday)

   Ta: krishna – Person name

   En: (Krishna)

9. Ta: chithirai ther thiruvilla - Entertainment event

   En: Chithirai boat festival

   Ta: chithirai – Month name

   En: Chithirai

3.3.1 Spell Variation

One of the major challenges in the web data is that we find different people spell the same entity with differently. With the advent of advancement of technology, people having access to internet have increased, and hence people creating content has increased. Some of the examples of spell variation are

lakshmi, latchmi, lacmi - lakshmi (Person)

nagarkoyil, nagarkovil - Nagarkovil(Place)

roca, roja - Rose (Person)

vaajpai, vaajpayee - vajpayee(Person)

sriperumpathur, ciriperumbathur - sriperumbudur(Place)

In the contemporary Tamil writings we observe that people avoid usage of Sanskrit alphabets and instead use a different alphabets, due to we find different spell variations. “ja”, “ha” are two common Sanskrit letters that are avoided in writing.

3.3.2 Name variations

A name variation is different from spell variation. Here the names of places or persons are changed due to either government notifications or localization by the public or by Anglicization. Here the names are different. For instance, “Chennai”, “Madras”. Some of the other examples are

thanjai , thanjavur

mumbai, bambay

thiruchi, thiruchirappalli

uthagai, ooty

kovai, coimbatore

3.4 Capitalization

In English and some other European languages Capitalization is considered as the important feature to identify proper noun. It plays a major role in NE identification. Unlike English capitalization concept is not found in Indian languages.

3.5 Noise in the data

The data obtained from the web or other online resources consists of lots of noise, which affects the processing of the text. Web pages are coded using mark-up programming languages, and the content is embedded inside these mark-up tags. The mark-up tags and other programming language codes add noise to the text. This codes and mark-up tags have to be removed from the web page to obtain clean text. At the outset, this cleaning seems to be a trivial task, but in some instances of web pages which do not follow proper set standards, removing the mark-up tags becomes difficult. Though this data cleaning task is not research problem, it’s an engineering issue, but still it’s very important to obtain clean data for further processing.

4 Experiments, Results and Discussions

The corpus is collected from various on-line tourism sites. The corpus consists of tourism and general domain. The corpus is divided into training and testing sets in the ratio of 70% and 30%. The corpus consists of 93K words. Table 1, shows the corpus statistics. The corpus is preprocessed for POS, and chunking. The preprocessing is done using automatic tools. Automatic POS (Arulmozhi and Sobha, 2006) and chunking (Vijay and Sobha, 2006) tools are built using machine learning techniques. The performance of these tools is in the range of 92% - 95%.

Corpus	No. Of Words	No. of NEs	Out-of Vocabulary NEs

Training	65022	12970	--
Testing	28270	5684	2035 (35.78%)
Total	**93292**	**18654**	

TABLE 1 – Corpus statistics

CRFs is trained with the training data to build the language model by using the features explained in section 2. Named entities in the Test data are identified and results are evaluated manually. Table 2, shows the evaluation results.

NE Type	Precision %	Recall %	F-score %
Person	82.52	81.30	81.92
Location	76.79	64.81	70.29
Numeric Expressions	69.76	81.16	75.03
Time Expressions	40.08	34.25	36.94
Other Categories	51.34	29.71	37.64
Over All	64.09	58.08	60.36

TABLE 2 – Base NE system evaluation results using CRF

SVM and CRF are both graphical learning methods and also we find that SVM is being used recently by the research community. Hence this has motivated us to choose SVM for comparison with CRF. We used Yamcha SVM classifier tool for NE classification and applied the same features in our base NE system using CRF. Table 3, shows the evaluation results for SVM.

NE Type	Precision %	Recall %	F-score %
Person	81.81	78.68	80.22
Location	75.43	59.51	66.53
Numeric Expressions	70.81	78.38	74.40
Time	38.82	29.44	33.49

Expressions			
Other Categories	55.75	27.79	37.09
Over All	64.52	54.76	58.34

TABLE 3 – Base NE system evaluation results using SVM

Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by Precision = tp/(tp + fp) where tp and fp are the numbers of true positive and false positive predictions for the considered class. Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. Recall = tp/(tp+fn) where tp and fn are the numbers of true positive and false negative predictions for the considered class, tp + fn is the total number of test examples of the considered class. The F-measure is the harmonic mean of precision and recall.

The above table clearly shows that in comparison with SVM, the recall score increases by 3% and f-measure range increases by 2% for Base NE system using CRF. The system performance is affected by the factors explained in section 3. Further, in this section, we explain the experiments performed to overcome some of the issues described in section 3.

To overcome the problem of agglutination for languages such as the one considered in this work, considering the root word instead of the actual word will be beneficial. We have performed experiments by taking the root words instead of the actual word. The evaluation results for the test data by taking the root words instead of the actual word is shown in the table 4. Use of root word has helped in improved identification of entities such as person names. Generally these entities occur with different case markers, by providing the root words for learning the system learns better.

NE Type	Precision %	Recall %	F-score %
Person	83.44	84.72	84.08
Location	72.57	72.26	72.42
Numeric Expressions	72.22	83.93	77.66
Time Expressions	37.63	35.95	36.77
Other Categories	54.30	43.74	48.41
Total	64.01	64.13	63.86

TABLE 4 – Evaluation results – when root words taken instead of actual words

Other major issue, we observe is, especially in the Time expressions, the special day NE category consists of either month name or person name NE embedded within it. This makes the NE identification system to tag both separately instead of one single entity. This makes the Time expressions NE identification harder. To overcome this issue of nested entities, use of post processing heuristic rules is reliable and convenient. This post processing has helped to improve Time expression detection. We have used linguistic and heuristic rules for post processing. The constraints we applied were based on the Part-of-speech tag , noun phrase chunk and a set of key-terms in tourism documents. We explain below some of the post-processing heuristic rules used in this work.

Rule 1: To attain the Entertainment tag

-2	NNP|NNPC|NNC|NN+B-NP=1	
-1	NNP|NN+B-NP=1	ENTERTAINMENT
0	urchavam(NN) =1	

The term “urchavam” indicates festivals held in temple. If the current token is “urchavam”, previous POS is a noun category(NNP,NNPC,NNC,NN) followed by beginning noun-phrase chunk(B-NP) and second previous POS from current token is a noun category followed by B-NP then the nested named entity is tagged as Entertainment.

-2	NNP|NNPC|NNC|NN=0	
-1	NNP|NNPC|NNC|NN=1	ENTERTAINMENT
0	urchavam(NN) =1	

If the current token is “urchavam”, previous POS is a noun category and second previous POS from current token is not a noun category then the nested named entity is tagged as Entertainment.

Rule 2: To solve the ambiguity exists between Person and Location tag.

-2	NNP|NNPC|NNC|NN+B-NP=1	
-1	NNP|NN+B-NP=1	PLACE
0	koyil|kovil|sannathi+NN =1	

The term “koyil|kovil|sannithi” means temple. If the current token is “koyil|kovil|sannathi”, previous POS is a noun category followed by beginning noun-phrase chunk and second previous POS from current token is a noun category followed by B-NP then the nested named entity is tagged as Location.

-2	NNP|NNPC|NNC|NN=0	
-1	NNP|NNPC+B-NP=1	PLACE
0	koyil+NN=1	

If the current token is “koyil|kovil|sannathi”, previous POS is a noun category followed by beginning noun-phrase chunk and second previous POS from current token is not a noun category then the nested named entity is tagged as Location.

Rule 3: To disambiguate the Person and Special-day tag.

-1 NNP|NNPC+PERSON=1 SPECIAL-DAY

0 jeyanthi+NN=1

The term “jeyanthi” denotes Birthday or day of celebration. If the current token is “jeyanthi”and the previous POS is a noun category then the nested named entity is tagged as Special-Day.

Rule 4: To get the Period tag

-1 QC=1 PERIOD

0 varutam|nAlY|ANtukalY|nAtkalY|mAwanworYum|nUrYrYANtu=1

The term “varutam|nAlY|ANtukalY|nAtkalY|mAwanworYum|nUrYrYANtu” indicates year| day|years|days|every month|century in English. If the current token is one among the key-terms specified in rule 4 and the previous POS specifies cardinal number then the nested named entity is tagged as Period.

Rule 5: To obtain the Distance tag.

-1 QC=1 DISTANCE

0 Ki.mi=1

The term “Ki.mi.” denotes Kilo meter in English. If the current token is “Ki.mi.” and the previous POS specifies cardinal number then the nested named entity is tagged as Distance.

Rule 6: To get the Time tag.

0 kalai|malai|iravu =1 TIME

1 QC =1

The term “kalai|malai|iravu” specifies morning|evening|night in English. If the current token is is one among the key-terms specified in rule 6 and the next POS specifies cardinal number then the nested named entity is tagged as Time.

-1 QC=TIME

0 mani

The term “mani” time in English. If the current token is “mani” and the previous POS specifies cardinal number then the nested named entity is tagged as Time.

In table 5, we show the evaluation results obtained after including the heuristic rules in the NE recognition system

NE Type	Precision %	Recall %	F-score %
Person	83.16	82.22	82.69
Location	74.20	64.48	69.00

Numeric Expressions	74.43	84.97	79.35
Time Expressions	66.40	56.29	60.93
Other Categories	57.20	33.53	42.28
Total	71.07	64.29	66.85

Table 5 - Evaluation results – after the application of heuristic rules (post processing) to overcome nested entity problem.

In the table 6, we show the evaluation results for the experiment where we have combined both the root word and the post processing. We find there is significant improvement in the results.

NE Type	Precision %	Recall %	F-score %
Person	84.37	86.07	85.22
Location	73.36	73.10	73.23
Numeric Expressions	76.74	87.28	81.62
Time Expressions	63.57	59.13	61.27
Other Categories	58.38	46.98	52.06
Total	71.28	70.51	70.68

TABLE 6 – Evaluation results – after the application of root word and heuristic rules (post processing)

To overcome the issue of spell variations and name variations, one approach can be use lexicons (or lists) containing equivalence NEs, for different variations. Other approach can be use of spell checker, by preprocessing the corpus prior with a spell checker engine and identify different variations, having threshold to consider equivalence classes.

Conclusion

In the paper we show that with the basic minimal features we obtain reasonable results. In the base system we don't make use of extra features and smoothing or post processing techniques. Then we have described about are various challenges encountered while developing the NE system. In the paper we discuss how these challenges can be overcome by using morph features, smoothing and other features, show improved results. Our NER system is developed efficiently so that it can handle different NE tags and unknown named entities. This is an ongoing work. We plan to further identify more semantic features from the corpus in the NE identification.

References

Arulmozhi, P. and Sobha, L. (2006). HMM-based Part of Speech Tagger for Relatively Free Word Order Language. *Advances in Natural Language Processing, Research in Computing Science Journal*, Mexico Volume18, pp. 37-48.

Bikel, D. M. Miller, S. Schwartz, R. Weischedel, R. (1997). Nymble: A high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing.* pp. 194201.

Borthwick, A. Sterling, J. Agichtein, E. and Grishman, R. (1998). Description of the MENE named Entity System. In *Seventh Machine Understanding Conference (MUC-7)*.

Chen, W. Zhang, Y. and Isahara, H. (2006). Chinese Named Entity Recognition with Conditional Random Fields. In *Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney. pp.118-121.

Ekbal, A. Bandyopadhyay, S. (2009). A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. *Linguistic Issues in Language Technology*, 2(1). pp.1-44.

Finkel, J. N. Grenager, T. and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005). pp. 363-370.

Finkel, J. Dingare, S. Nguyen, H. Nissim, M. Sinclair, G. and Manning, C. (2004). Exploiting Context for Biomedical Entity Recognition: from Syntax to the Web. In *Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, *(NLPBA)*, Geneva, Switzerland.

Gali, K. Surana, H. Vaidya, A. Shishtla, P. Sharma, D. M. (2008). Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition. In *Workshop on NER for South and South East Asian Languages, IJCNLP-08*, Hyderabad, India.

Klinger, R. Christoph, M. Friedrich, Fluck, J. Hofmann-Apitius, M. (2007). Named Entity Recognition with Combinations of Conditional Random Fields. In *2Biocreative Challenge Evaluation Workshop, CNIO*, Madrid,Spain. pp. 89-92

Krupka, G. R. and Hausman, K. (1998). Iso Quest Inc: Description of the NetOwl Text Extraction System as used for MUC-7. In *Seventh Machine Understanding Conference (MUC 7)*.

Kudo,T. (2005). CRF++, an open source toolkit for CRF, *http://crfpp.sourceforge.net*

Kudo,T. (2005). YamCha: Yet Another Multipurpose CHunk Annotator, an open source toolkit for CRF, *http://chasen.org/~taku/software/yamcha/*

Kumar, K. N. Santosh, G. S. K. Varma, V. (2011). A Language-Independent Approach to Identify the Named Entities in under-resourced languages and Clustering Multilingual Documents. In *International Conference on Multilingual and Multimodal Information Access Evaluation*, University of Amsterdam, Netherlands.

Lafferty, J. McCallum, A. Pereira, F. (2001). Conditional Random Fields for segmenting and

labeling sequence data. In *ICML-01*, pp. 282-289.

Loinaz, I.A. Uriarte, O. A. Ramos, N. E. Castro, M. I. F. D (2006). Lessons from the Development of Named Entity Recognizer for Basque. *Natural Language Processing*, 36. pp. 25 – 37.

McCallum, A. and Li, W. (2003). Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In *Seventh Conference on Natural Language Learning (CoNLL).*

Nadeau, David and Sekine, S. (2007) A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1). pp.3–26.

Niu, C. Li, W. Ding, J. Srihari, R. K. (2003). Bootstrapping for Named Entity Tagging using Concept-based Seeds. In *HLT-NAACL'03*, Companion Volume, Edmonton, AT. pp.73-75.

Pandian, S. Lakshmana, Geetha, T. V. and Krishna. (2007). Named Entity Recognition in Tamil using Context-cues and the E-M algorithm. In *the Proceedings of the 3rd Indian International Conference on Artificial Intelligence*, Pune, India. pp. 1951 -1958.

Sasidhar, B., Yohan, P.M., Babu, V.A., Govarhan, A.(2011). A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu. J. *International Journal of Computer Science Issues*, Volume. 8, pp. 1694-0814 .

Settles B. (2004).Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA),* Geneva, Switzerland. pp.104-107.

Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In *HLTNAACL 03*. pp.213-220

Sobha, L., Vijay Sundar Ram. R. (2006). "Noun Phrase Chunker for Tamil", In Proceedings of Symposium on Modeling and Shallow Parsing of Indian Languages, Indian Institute of Technology, Mumbai, pp 194-198.

Srihari, R.K. Niu, C. Yu, L. (2000). A Hybrid Approach for Named Entity Recognition in Indian Languages. In *6th Applied Natural Language Conference*, pp. 247-254

Gupta, S. and Bhattacharyya, P. (2010). Think globally, apply locally: using distributional characteristics for Hindi named entity identification. In *2010 Named Entities Workshop, Association for Computational Linguistics Stroudsburg, PA, USA*

Vijayakrishna, R. and Sobha, L. (2008). Domain focused Named Entity for Tamil using Conditional Random Fields. In *IJNLP-08 workshop on NER for South and South East Asian Languages*, Hyderabad, India. pp. 59-66