Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 159 – 166

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Evaluation of Named Entity features for Punjabi Language

Amandeep Kaur[a,*], Gurpreet Singh Josan[b]

[a]University College of Engineering, Punjabi University, Patiala –147001, India
[b]Department of Computer Science, Punjabi University, Patiala –147001, India

Abstract

Named entity recognition is a task to identify and classify the words in the given text to some predetermined categories like Organization, Location, Time, Number, Person etc. In this paper, NER system for Punjabi language has been evaluated on various combinations of features including context word window feature of 3, 5 and 7, various digit features, infrequent word and length of word features.After evaluation it has been found that the feature set comprising of word window 5, digit features, Infrequent word and Length of word feature has confirmed the highest f-score value of 87.46% with Precision and Recall values of 90.99% and 84.19% respectively. It has been realized that the feature set consisting of all language independent features give better results with word window 5 as compared to word window 3 and 7.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

*Keywords:*Named Entity Recognition; Named Entities; Punjabi language; Digit features; Length of word; Infrequent word

1. Introduction

Named Entities (NE) are phrases that represent Organization, Location, Time, Number, Person etc. in a given text. Named Entity Recognition (NER) is a language computational task in which every word in a document is

* Corresponding author. Tel.: +91-81465-87682
E-mail address: amandhillon83@yahoo.co.in

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.007

classified as falling into some predetermined classes like location, organization, time, person, percentage, monetary value and others (none-of-the-above).[1]NER involves two tasks: identification of NEs and classification of NEs into different classes such as organization, location, person etc. Formally, NER can be defined as a labeling taskthat labels an input sequence of words$W_1^n = w_1 \ldots\ldots w_n$, with a label sequence $L_1^n = l_1 \ldots\ldots l_n$,where label $l_i$given to a word is either a predetermined classes for named entities orit is other. Moreover, NEs are special words which are not defined under the grammatical rules of a language.

First introduced at Sixth Message Understanding Conference-6 (MUC-6), NER task is as an important sub-task of Information Extraction (IE) [6]. After that proper identification and classification of NEs has attracted the attention of Natural Language Processing (NLP) researchers. Information Retrieval and Extraction (IREX) program [13], Automatic Content Extraction (ACE)[18] program, Conferences on Natural Language Learning 2002 and 2003(CoNLL 2002 and 2003)[12,11] have large contribution inemergence of NER. NER plays an important role in various Natural Language Processing (NLP) applications like Text Summarization,Question Answering systems, Information Extraction and Retrieval,Machine Translation etc.

Some foreign languages which include English, German, Spanish, Arabic etc. have been deeply explored in the area of NER with high accuracy. Although in recent years NER research for Indian languages has also taken pace but still the accuracy of these NER systems is not comparable with English and other foreign languages. The Capitalization feature which has been proved to be very useful in English language is not applicable to Indian languages as the later are case-insensitive. The NERSSEAL Workshop initiated NER research for ILs as a shared task. This workshop focused on Hindi, Oriya, Bengali, Urdu, and Telugu. The tagset defined for this task consists of 12 tags namely PERSON, ORGANIZATION, LOCATION, DESIGNATION, ABBREVIATION, TITLE-PERSON, BRAND, TITLE-OBJECT, TIME, MEASURE, NUMBER, TERMS[16].

Like many other Indian languages, Punjabi Language was nowhere in the scene during NERSSEAL workshop. Punjabi has a very old and rich literary history and is also going towards fast technological developments.Punjabi text is availablein electronic form in both Gurmukhi (for Indian Punjabi language) and Shahmukhi (for Pakistani Punjabi language) scripts.Punjabi, the official language of Punjab state in India, is written using Gurmukhi script. With the availability of Gurmukhi text in electronic form, researcher starts developing systems like Parts of Speech tagging, Machine Translation, Question Answering Systems etc. for Punjabi language. The accuracy of all these systems is not comparable to their counter-parts in other well researched languages like English mainly due to lack of supporting systems. One such supporting system is Named Entity Recognizer. Non availability of NER system in Punjabi language isone of the major hurdles in these research activities and thus our motive to pursue research in this direction. The NER research work in Punjabi language was initiated with 12 tags using Conditional Random Fields approach as presented in [8].

In this paper various experiments, performed using different combinations of named entity features have been discussed. The paper is divided into six parts. Next section describes various NER approaches. In Section 3 related work is discussed. Section 4 describes work regarding NER in Punjabi Language. Section 5comprises of various experiments and their results. Finally in Section 6 work has been concluded.

2. NER Approaches

According to[9], the methods used for automatic identification and classification of named entities, are classified into three categories.

- Rule-based NER – These systems focuses on extracting NEs using hand-made rules. Most of the earlier studies were based on hand-crafted rules.
- Machine Learning-based NER - These systems find patterns and their relationships in the given text to prepare a model using a machine learning approach. These systems can be further classified as:
  Supervised machine learning model –This approach constructs a statistical model based upon atagged training data.Various approachesare Hidden Markov Model (HMM), Conditional Random Fields (CRF), Maximum Entropy (ME), Decision Trees(DT), and Support Vector Machines (SVM).
  Unsupervised machine learning model-In this approach an unsupervised model learns without any feedback. The aimis to construct representations from data which are further used in identification and classification of NEs.
- Hybrid NER – These systems are the combination of both of the above systems.

3. Related Work

The Third International Joint Conference on Natural Language Processing(IJCNLP – 08) workshop on NER for South and South East Asian Languages (NERSSEAL), held in 2008 at IIIT Hyderabad, was an important initiativein the direction of NER research for Indian languages with focus on Bengali, Oriya, Hindi, Telugu and Urdu languages. 12 research papers were presentedin this workshop and out of which four were related to shared task. All these research papers were based on 12 tags NE tagset defined for NERSSEAL[16].

Sujan Kumar Saha in[10] described a system which combined Maximum Entropy model with language specific rules and gazetteer lists. Language specific rules were prepared for Bengali and Hindi only. For the remaining languages the systemswere developed usingMaximum Entropy approach. The system recognized twelve classes of Named Entities.The reported f-value for Hindi is 65.13%, for Bengali is 65.96%. The system also works for Telugu, Oriya and Urdu. The reported f-scorefor these languages are 44.65%, 18.74%, and 35.47%respectively.

The work regarding Telugu language is mentioned in[15]. They used Conditional Random Fields (CRF) approach for recognizing named entities using various language dependent and language independent features and reported precision64.07%, recall 34.57% and F-measure of 44.91%.

The work reportedin[3]is about the development of a NER system for Bengali language. The authors used Support Vector Machine (SVM) approach. For this experiment an overall average Recall is 94.3%, Precision is 89.4% and F-Score is claimed to be91.8%.

In[4] author reported the development of CRF based NER system. This system used both language independent as well as dependent features. Evaluation results have demonstrated the highest F-Score of 59.39% for Bengali, 33.12% for Hindi, 28.71% for Oriya and 35.52% for Telugu.

The work in[5] is also based upon Conditional Random Field approach. In this work authors have combined machine learning techniques with language specific heuristics. They have reported F-measure of 43.46%, 40.63%, 50.06%, 39.04% and 40.94% for Urdu, Bengali, Hindi, Oriya and Telugu respectively.

Authors presented NER work for Telugu language in[17]. They developed a system for Telugu language which is based upon Conditional Random Fields approach. The system was tested on differentdata sets for identifying person, location and organization names. F-score between 80% and 97% was obtained in various experiments.

A three-stage approach for named-entity identification for Bangla language has been discussed in[2]. The stages are based on the use of NE dictionary, rules for NE and left-right co-occurrence statistics. Named Entities are identified onlyand are not classified. Overall F-score of 89.51% was reported by the authors.

The initial research in NER for Punjabi language is presented in[8]. In this work, a small annotated corpusand few gazetteers were manually created using NERSSEAL tagset of 12 tags. Corpus was also manually tagged using coarse grained Parts-of-Speech (POS) tagset of 9 tags.

4. NER in Punjabi Language

Punjabi is a descendent of Indo-Aryan language.It is spoken by people of the Punjab in India and Pakistan Punjabi is the official language of the Indian state of Punjab and also one of the official languages of Delhi. According to the Ethnologue 2005[19] estimate there are more than 88 million speakers of the Punjabi language and ranked 20th among the languages spoken in world.

It has been observed that vast amount of information about Punjabi language is available online, but this information is not present in a proper format which could be used to benefit the local users. Punjabi, like other Indian languages, is still lacking in the availability of resourcesin the required measure. Web sources for various gazetteer lists are not available in Punjabi. This leads to little attention for Punjabi Language in Natural Language Processing (NLP) tasks especially in the area of Named Entity Recognition.

NER is an important NLP task which has not been deeply explored for Punjabi Language and thus the motivation for this research. A small initiative in this direction was taken in 2009 as discussed in[8]. A small annotated corpus and gazetteers were manually created. Corpus was also manually tagged using coarse grained Parts of Speech (POS) tagset of 9 tags. Further extending this work, it was decided to create more exhaustive and improved resources for developing a Punjabi NER system.

Although various NER approaches have been identified and reported by authors for developing NER systems but Machine Learning approachesbenefit the most in this area. The efficiency of machine learning techniques is highly dependent on large corpora annotated with named entities.So, it was decided to prepare an NE tagged annotated

corpus for Punjabi language first, as none was available in the literature. In corpus annotation task two important challenges were faced: deciding NE Tagset (annotation categories) and annotation guidelines. It was soon realized that without clearly defined annotation guidelines and tagset, it is very difficult for the annotators to perform manual and/or automatic annotation. Even if the proper annotation guidelines and tagset are provided before annotation task, still there is high probability of amendments. As the annotation task progresses, more ambiguous and confusing cases are identified that do not fit the given tagset and annotation guidelines thus leads to inconsistency in the corpus. In[7], various tagset design issues and problematic cases faced during the annotation of corpus of Punjabi NER were discussed and accordingly proposed additional tags to be used for NER task in Punjabi language. In order to develop the tagset, Extended Named Entity hierarchy provided in[14], CoNLL 2002 and 2003 tagset and 12 tags of NERSSEAL were referred.

For developing the annotated corpus, two online Punjabi newspapers Ajitweekly.com[20] and Ajitjalandhar.com[21] were used. Ajitjalandhar was found to be more useful as it also provides web archive as well as it is the most popular newspaper in Punjab region of India. E-news articles corresponding to Sports, Business, Entertainment, Health, Religion, and General news from State, National and International fields were collected.

4.1. Named Entity Tagset

In NERSSEAL Workshop the tagset defined consisted of 12 tags as listed in table 1.[16]

Table 1 Named Entity Tagset

Name	Tag	Examples
Person	NEP	ਰਣਜੀਤ [Ranjit]
Location	NEL	ਪੰਜਾਬ [Punjab]
Organization	NEO	ਕਾਂਗਰਸ [Congress]
Facility	NFAC	ਅਪੋਲੋਹਸਪਤਾਲ[Apollo Hospital]
Event	NEVE	ਓਲੰਪਿਕਖੇਡਾਂ[Olympics]
Relationship	NREL	ਭਰਾ [Brother]
Time	NETI	ਦਸਸਾਲ [10 Years]
Date	NEDA	ਜੂਨ 2013 [11June 2013]
Designation	NED	ਮੰਤਰੀ [Minister]
Title-Person	NETP	ਸੰਤ [Saint]
Number	NEN	ਇੱਕ [One]
Measure	NEM	10 ਪ੍ਰਤੀਸ਼ਤ [10 %]
Abbreviation	NEA	ਆਈ.ਪੀ.ਐਲ [IPL]
Artifact	NART	ਪੰਜਾਬੀ [Punjabi language]
Other (Not an NE)	O	

In[8], for evaluation of NER in Punjabi language, NERSSEAL tagset of 12 tags was used. During this work various ambiguous and problematic cases were realized. Various tagset design issues and author's recommendations about additional tags were presented and finally a tagset of 14 tags have been proposed in[7]which will be used in the current research work. Although Extended Named Entity hierarchy provides more than 100 tags but we have opted a limited tagset. Keeping in view the resource scarcity issue and the initial stage of NER research in Punjabi language, we proposed an NE tagset of 14 tags namely PERSON, ORGANIZATION, LOCATION, FACILITY, EVENT, RELATIONSHIP, TIME, DATE, DESIGNATION, NUMBER, TITLE-PERSON, MEASURE, ABBREVIATION and ARTIFACT.

The corpus has been annotated using the tagset mentioned in Table 1. For each Named entity a tag has been specified. For instance, for annotating single person name in the data, the tag NEP is used that denotes Named Entity Person.

In order to tag a multiple entity which consists of more than one word, IOB tagging scheme is used. For instance, for annotating First name of a person, the tag B-NEP (Beginning of Named Entity Person) is used and for annotating Last name and middle name, the tag I-NEP (Intermediate of Named Entity Person) is used.

4.2. Architecture of Punjabi NER system

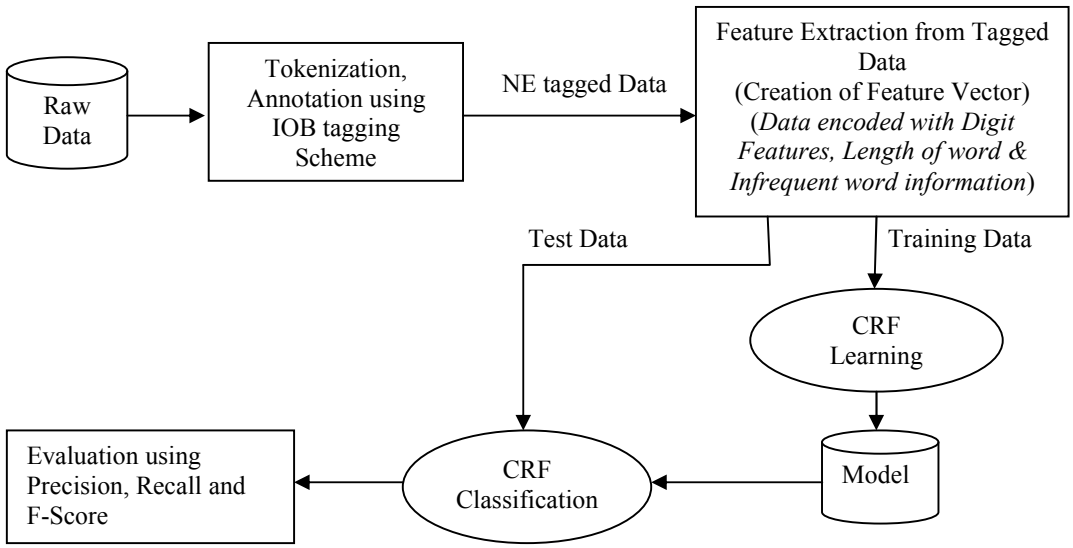

Fig.1 Architecture of Punjabi NER system

Fig. 1 shows a proposed architecture of Punjabi NER system in which raw data is tokenized into column format with one word per line and sentences separated with a single blank line. This data is annotated using Named entity tagset and IOB tagging scheme. The NE tagged data is fed to the Feature extractor module which encodes each word with 11 binary features (Feature vector).In CRF learning module, training data and the selected features from training data are used to create a model for the system. This model is further used by CRF classification module to classify NEs in test data. In last phase, three parameters viz. Precision, Recall and F-score are used to evaluate the system.

5. Named Entity Features for Punjabi language

Named entity features are very helpful in identification and classification of NEs.These features can be language independent or language dependent. As the name implies, Language independent features are applicable to any language without having deep understanding of that language but Language dependent features require core understanding and knowledge of the respective language. Various Language independent features are Word prefixes and suffixes, context words, first word, infrequent word, word length, digit informationetc. Features like set of known prefixes and suffixes of the words, clue words whichhelpin the prediction of various Named Entities like

person, organization, facility, events etc., various gazetteer lists, and Parts of Speech information of the given word are language dependent.

In this work, we have considered various combinations of named entity features in order to identify the best feature set for NER in Punjabi language. Following features have been considered:

- Context word feature: Context words are neighbors of candidate word. As context words can be used to identify NEs, we use context word as a feature. In our work we have considered word windows of size 3, 5 and 7.
- Digit features: It is a binary valued feature used for the recognition of expressions which can be NEs like time, measurement, date and numbers etc. Various digit patterns have been defined based on the occurrence of digits in the word like
  Ctndg [word consists of digits],
  Ctnfourdg[word is of four contiguous digits],
  Ctntwodg[word is of two contiguous digits],
  Ctndg,[word comprised of digits followed by comma],
  Ctndg.[word comprised of digits followed by period],
  Ctndg/[word comprised of digits followed by slash],
  Ctndg-[word comprised of digits followed by hypen],
  Ctndg%[word comprised of digits followed by percent],
  Ctndgquotes[word consists of digits and quotes].
- Infrequent word: Frequent words are rarely NEs. This can be used as a feature. From the training corpus, a list of infrequent words was prepared. To represent this feature, we use binary value (0 or 1) based upon whether a word appears in infrequent wordlist or not.
- Word Length: If word length is small it is unlikely to be Named Entity. This binary valued feature checks the length of candidate word. If it is more than three then feature value is 1, otherwise 0.

6. Experiments and Results

C++ based OpenNLP CRF++ package[22], which isbased on Conditional Random Fields (CRFs) approach for labelling sequential datahas been used for developing Punjabi language based NER system. This NER system was trained on training data of 1,70,000 words and tested with test data of 30,000 words.For every word in the tagged corpus, a feature vector is extracted. Thus the training corpus consists of words, feature vectors and answer tags. Various CRF Models are built using training data and different feature templates.

Various experiments have been conducted using different combinations of features with the aim to identify the optimal feature set.From experimental analysis, it was found that the following feature set F of named entity features gives the best result for the test data.

F = [context word feature of word window 5, all digit features, Infrequent word feature]

For evaluating results, the standard evaluation parameters Precision, Recall and F-Score have been used. The results for test data are presented in Table 2.From the results, it has been realized that context word window of size 3,5 and 7 gives similar results without much difference in F-score values (1^{st}- 3^{rd} rows). Inclusion of digit features to context word window 3, 5 and 7 further increases the f-score value by 1.29%, 0.38%, and 0.46% respectively (4^{th} – 6^{th} rows). The addition of 'Infrequent word' feature to context word window 3 has reduced the f-score by 0.58% (10^{th} row) whereas addition of 'Infrequent' word feature to word window 5 and 7 has improved the f-scores values by 0.17% and 0.25% respectively(11^{th} and 12^{th} row). The use of length feature along with digit features seems to be quite useful to word window 3, 5 and 7 as it has improved the f-score value to 87.40%, 87.38% and 87.25% respectively (16^{th} to 18^{th} row). The addition of 'Infrequent' word feature to above feature set reduces the f-score value in word window 3 but increased the f-score values of word window 5 and 7(19^{th} to 21^{st} row). So, the feature set with highest f-score value of 87.46% comprises of word window 5, digit features, 'Infrequent' word and Length of word feature. The second highest feature set that has given comparable f-score value of 87.40 % comprises of word window 3, digit features and length feature. It has been found that feature set combining all language independent features with word window 5 gives highest results.

Table 2 Evaluation Results

Features	Precision(%)	Recall(%)	F-Score(%)
F1 = wi-1,wi,wi+1	89.70	82.78	86.10
F2= wi-2, wi-1,wi,wi+1,wi+2	89.74	84.04	86.80
F3= wi-3,wi-2, wi-1,wi,wi+1,wi+2,wi+3	89.21	84.06	86.56
F4= F1 + digit feature	90.55	84.44	87.39
F5= F2 + digit feature	89.84	84.67	87.18
F6 = F3 + digit feature	89.54	84.63	87.02
F7= F1+ infrequent word	91.18	82.85	86.82
F8= F2 + infrequent word	90.93	83.09	86.84
F9= F3 + infrequent word	90.72	84.04	86.71
F10 = F4 + infrequent word	90.87	83.10	86.81
F11 = F5 + infrequent word	90.96	84.02	87.35
F12 = F6 + infrequent word	91.05	83.78	87.27
F13 = F1 + length	89.98	84.28	87.04
F14 = F2 + length	89.73	84.21	86.88
F15 = F3 + length	89.58	84.08	86.74
F16 = F4 + length	89.99	84.94	87.40
F17 = F5 + length	89.92	84.97	87.38
F18 = F6 + length	89.79	84.84	87.25
F19 = F10 + length	90.76	83.30	86.87
F20 = F11+ length	**90.99**	**84.19**	**87.46**
F21 = F12 + length	90.76	84.23	87.37
wi – current word			
wi-1 – previous word			
wi+1 – next word			
wi-2 – 2nd previous word from current word			
wi+2 – 2nd next word from current word			
wi-3 – 3rd previous word from current word			
wi+3 – 3rd next word from current word			

7. Conclusion

This paper presented the evaluation of various Named Entity features that are quite helpful in recognition of named entities in Punjabi language. These features include context word window of size 3, 5 and 7, various digit features, Infrequent word feature and Length of word feature. Experiments have been conducted on various combinations of these features using Conditional Random Fields approach. The feature set comprising of word window 5, digit features, Infrequent word and Length of word features has confirmed the highest f-score value of 87.46%. The second highest F-score value of 87.40% was found on evaluating feature set containing context word window 3, digit features and Length features. It has been realized from the evaluation that the feature set combining all language independent features with word window 5 gives highest results.

References

1. Borthwick, A., Maximum Entropy Approach to Named Entity Recognition, Ph.D. dissertation, Computer Sci. Dept., New York Univ., New York, USA,1999.

2. Chaudhuri, B. B. and Bhattacharya, S., An Experiment on Automatic Detection of Named Entities in Bangla,IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 75-82.
3. Ekbal, A. and Bandyopadhyay, S., Bengali Named Entity Recognition using Support Vector Machine,IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 51–58.
4. Ekbal, A., Haque, R., Das, A., Poka V. and Bandyopadhyay, S., Language Independent Named Entity Recognition in Indian Languages, IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 33–40.
5. Gali, K., Surana, H., Vaidya, A., Shishtla, P. and Sharma, D. M., Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition, IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 25-32.
6. Grishman, R. and Sundheim B., Message Understanding Conference - 6: A Brief History, 16th International Conference on Computational Linguistics (COLING), 1996, p 466 – 471.
7. Kaur, A. and Josan, G., Improved Named Entity Tagset for Punjabi Language,2014 RAECS, 2014, doi 10.1109/RAECS.2014.6799638.
8. Kaur, A., Josan, G. and Kaur, J., Named Entity Recognition For Punjabi: A Conditional Random Field Approach, ICON-2009: 7th International Conference on Natural Language Processing, 2009,p 277-282.
9. Mansouri, A., Suriani Affendey, L. and Mamat, A., Named Entity Recognition Approaches, International Journal of Computer Science and Network Security, 2008,p 339-344.
10. Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S. and Mitra, P., A Hybrid Approach for Named Entity Recognition in Indian Language,IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 17-24.
11. Sang, E. F. T. K. and Meulder, F. D., Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,7th Conference on Natural Language Learning CoNLL-2003, 2003.
12. Sang, E. F. T. K.,Introduction to the CoNLL- 2002 shared task: Language-independent named entity recognition,6th Workshop on Computational Language Learning, CoNLL-2002, 2002.
13. Sekine, S.and Ishara, H., IREX: IR & IE evaluation project in Japanese, 2nd International Conference on Language Resources and Evaluation, 2000.
14. Sekine, S., Sudo, K. and Nobata, C., Extended Named Entity Hierarchy, 3rd International Conference on Language Resources and Evaluation, LREC, 2002.
15. Shishtla, P. M., Gali, K., Pingali P. and Varma, V., Experiments in Telugu NER: A Conditional Random Field Approach, IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008, p 105-110.
16. Singh, A. K., Named Entity Recognition for South and South East Asian Languages: Taking Stock,IJCNLP-08 Workshop on NER for South and South East Asian Languages,2008, p 5–16.
17. Srikanth, P. and Murthy, K. N., Named Entity Recognition for Telugu,IJCNLP-08 Workshop on NER for South and South East Asian Languages, 2008,p 41-50.
18. ACE website,http://www.itl.nist.gov/iaui/894.01/tests/ace/2000 /doc/ace-tides00/sld007.html
19. Ethnologue 2005website, http://en.wikipedia.org/wiki/Ethnologue
20. Ajitweeklywebsite ,http://www.ajitweekly.com
21. Ajitjalandharwebsite,http://www.ajitjalandhar.com
22. CRF++ packagewebsite, http://crfpp.sourceforge.net