Named Entity Recognition System for Urdu

UmrinderPal Singh[1,1], Vishal Goyal[2,1] Gurpreet Singh Lehal[3,1]
(1) Department of Computer Science, Punjabi University Patiala
umrinderpal@gmail.com[1], vishal.pup@gmail.com[2], gslehal@gmail.com[3]

ABSTRACT

Named Entity Recognition (NER) is a task which helps in finding out Persons name, Location names, Brand names, Abbreviations, Date, Time etc and classifies them into predefined different categories. NER plays a major role in various Natural Language Processing (NLP) fields like Information Extraction, Machine Translations and Question Answering. This paper describes the problems of NER in the context of Urdu Language and provides relevant solutions. The system is developed to tag thirteen different Named Entities (NE), twelve NE proposed by IJCNLP-08 and Izaafats. We have used the Rule Based approach and developed the various rules to extract the Named Entities in the given Urdu text.

Keywords: NER, Urdu NER, Urdu, Rule Based

Proceedings of COLING 2012: Technical Papers, pages 2507–2518,
COLING 2012, Mumbai, December 2012.

1. Introduction

Named Entity Recognition (NER) is a subtask of Information Extraction (IE). NER extracts and classifies the true Named Entities in text. NER system is widely used in different tasks of Natural Language Processing (NLP) and in many commercial applications on internet like Search Engine. Accuracy of NER system is directly reflected in NLP applications. So, accurate working of NER system is very important. NER system can be used for one's personal interest like company manager wants to know all the names involved in specific text document.

S.No:	**Tag Name**	**Example**
1	Person Name	ارشد(Arshad)
2	Location	پٹیالا (Patiala)
3	Organizations	رلاینس (Reliance)
4	Terms	سپنڈلاٹس (Spondylitis)
5	Designation	وزیر اعظم (President)
6	Title Person	جناب (Mr.)
7	Title Object	ہندوستان ٹائمز (Hindustan Time)
8	Brand	سیمسنگ (Samsung)
9	Measure	سال 10 (10 Years)
10	Number	ایک، دو (One, Two)
11	Date/Time	اکتوبر 12 (12 October)
12	Abbreviation	بی بی سی (BBC)

TABLE 1- Different Named Entity Tags

Named Entities mentioned above were proposed at IJCNLP-08 workshop [17]. Named Entities can be domain specifics like NER system to identify entities in scientific data.

The NER system can be developed using three approaches, 'Rule-Based', 'Machine Learning' (HMM, SVM, CRF, Decision Tree) and 'Hybrid' approach. The Rule-Based system is difficult to develop as one should know the language and grammar rules. These kinds of systems are domain specific. Machine learning approach provides different Statistical NLP tools to train NER system. Statistical tools provide fast way to develop NER system but the accuracy of the system is dependent on annotated training data. For greater accuracy we need to train the NER system with large amount of annotated data. Hybrid approach is a combination of both Rule Base and Statistical based.

2. Related Work

NER system came in focus during the sixth Message Understanding Conference (MUC-6) [6]. After that many NER systems were developed. Most of these systems were developed for European languages and all systems were highly accurate. For south Asian languages, NER systems yet in developing phase. IJCNLP-08 workshop played a major role in development of NER Systems for Indian languages. This Workshop focused on five languages i.e. Hindi, Bengali, Oriya, Telugu and Urdu. All the systems were developed using Statistical approaches or Hybrid approach. Hybrid NER system for five languages was developed by (Sujan Kumar saha et al. 2008) [2]. Rules were developed only for Hindi and Bengali. The system was developed using MaxEnt model. Accuracy for Urdu was Maximal, Nested and lexical were 27.79, 28.59 and 35.47 respectively.

Karhik Gali et al.2008 [18] had developed the system for five languages Telugu, Hindi, Bengali, Urdu and Oriya. The system was developed using CRF based machine learning model. This system also used some heuristic rules. The system was specific for Telugu and Hindi. Accuracy for Urdu was Maximal, Nested and lexical, were 39.86, 39.01 and 43.46 respectively. Asif Ekbal et al 2008 [1] had developed the system using CRF Machine learning approach. The system was trained for Bengali, Hindi, Telugu, Oriya and Urdu. The system also used language dependent and language independent rules. Accuracy of the system for Urdu was Maximal nested and lexical, were 30.35, 28.55 and 35.52 respectively. Praveen Kumar P et al 2008[3] had developed the system for Hindi, Bengali, Oriya, Telugu and Urdu languages. The system was developed using Hybrid approach which was a combination of CRF and HMM models. Accuracy of the system for Urdu by using CRF, Maximal Nested and Lexical were 33.17, 31.78 and 38.25 respectively and by using HMM 34.48,36.83 and 44.73 respectively . Amit Goyal 2008[10] had developed the system using CRF machine learning model for Hindi language. Accuracy of the system was 58.85. Shilpi Srivastava et al. 2011 had developed the NER system for Hindi language based on CRF and MaxEnt models of Machine Learning approach and rules were developed for Hindi language. The system used voting method to improve the accuracy. Accuracy of the system was 82.95. Kashif Riaz et al. 2010[4] had developed the system using the Rule Based approach. Rules were developed for Person name, Location, Date, Numbers, Organizations and Person's designations tags. The system used very small gazetteer for person names and locations. Recall of the system was 90.7%, precision 91.5 and F1-measure was 91.1%. which was better than all the NER systems developed in IJCNLP-2008 workshop for Urdu? We can see that sufficient work was not done for Urdu NER system and work which is available does not show satisfactory results. Only Kashif Riaz's work shows good results.

3. Approaches to NER

3.1 Rule Based approach: Rules are developed to identify NE in text. This approach takes much time in development and one should have good knowledge of target language. Heuristic based rules are used to identify tags and these rules are language specific. Good rules always yield good results. Development of these kinds of systems is always a time consuming task.

3.2 Statistical approach: Statistical approach is also known as Machine Learning approach. This is a fast way to develop a NER system. The system is trained using annotated training data set in specified format. Accuracy of statistical approach is dependent upon the training data. So, we always train the system with a large set of annotated data. Various Machine Learning models like HMM, CRF, MaxEnt, are used for NER system.

3.3 Hybrid system: Hybrid system is combination of Rule Based approach and Statistical approach. To develop the Hybrid system we use Statistical tools as well as linguistic rules. Combinations of both approaches make a system more accurate and efficient.

4. Issues in Urdu NER System

No capitalization: Urdu and other Asian Languages do not have concept of capitalization. In European language like English this feature is widely used to recognize Named Entity in text because all the names in text always begin with capital letter. Absence of capitalization feature makes the NER task hard for Urdu language.

Ambiguous Name: Urdu language has lots of ambiguous names that can be used as proper noun as well as common noun. Main challenge of any NER system is to separate or extract proper noun in place of common noun. Example: برکت (barkat) or سلامت (slaamat) can be the name of a person or it can be used as common noun.

Spelling variations: Lack of standardization in Urdu language can be seen in spelling as well. There are different spellings that can be used for same word. Like word Hospital can be written in two ways in Urdu اسپتال/ہسپتال (hasptaal/asptaal) which makes the task difficult for NER. We are unable to collect the standard spelling of foreign language. Example: انسٹیچیوٹ/انسٹیٹوٹ (institutes/instichutes).

Non-availability of resources: Language Resources are must for any approach whether it is Rule Based or Statistical. There is no large gazetteer and annotated data available for Urdu language.

5. Why Rule Based Approach

Rule Based approach is time consuming task to develop any NER system. Rule based approach is used only when you know the target language well and have sufficient knowledge about the linguistic rules like knowledge of grammar. The system developed using Rule Based approach always yields the good results. On the another hand, Statistical approach which provide us with many Statistical tools, to develop NER system like HMM, CRF, SVM, MaxEnt etc, with the help of these tools development process of the system is rapid as compared to Rule Based approach. We have studied that in IJCNLP-2008 that all the NER systems were developed using different Statistical approaches. But none of the system provides good results for Urdu text because annotated data provided by the workshop is only 36000 Urdu tokens which is not sufficient to train Urdu NER system. New Statistical techniques like CRF not perform well for Urdu. Absence of any large Urdu gazetteers is also one of the reasons for low accuracy. These kinds of gazetteers boost the accuracy of Statistical approaches. Rule Based approach used by (Kashif Riaz 2010[4]) for Urdu language shows good results. Rules are used to identify six tags. Workshop on NER system by IJCNLP 2008 focused on twelve NER tags. By studying various research papers we concluded that we should follow Rule Based approach though it is a time consuming approach but this approach will give us promising results as we have seen in Kashif Riaz's[4] system. We did not try Hybrid Approach because the absence of large annotated corpus for Statistical part. We have tried to develop different rules for all 12 NE tags which are used in IJCNLP-08 workshop.

6. Rule Based Model

Following Rules are used to identify different tags in Urdu text.

1. Rules are applied to identify date and time tags. These kinds of tags are easily identified by Regular Expressions that are created for specific patterns like 01.01.2012 or 01/01/2012 and time is also identified as 11:20 or 01:22. The system is able to identify the date like 01 May 2012 or 01 May and year 2012.

2. Suffix matching is used to identify various locations and types of names and terms. In Urdu language and other south Asian languages, there are many location names that end with 'pur' (Kishanpur, Rampur), 'stan' (Pakistan, Hindustan) 'ghar' (Chandighar, Ramghar), 'nagar'

(Sonagar) and words that end with 'abad' (Fridabad, Hardabad). Suffix matching is also used for persons name, terms and org Like, person name that ends with 'dev' (Ramdev, Shamdev), 'das' (Sumitvadas, Charndas). Terms that ends with 'logy' (biology), person's last name ends with 'brown' or 'wood' then we can identify them as person's name or it may be organization like Hollywood and Bollywood.

3. The system uses gazetteer of most common person names to identify 'Person Names tag'. The system is able to tag words of maximum three lengths as one Named Entity like محمد شفیق تھند (mahmad saphīk thindh). For person names we have collected 4500 Urdu names and 1500 Hindu person names.

4. We have collected the surname of Muslim and Hindu religions. With the help of surnames the system is able to identify his/her first name, like surname (khān)خان helps the system to check one word before the surname which may be the first name of person like (shāhrukh khān) شاہ رُخ خان

5. Title person and Designations helps the system to identify person name like Title Person وزیر اعظم (vajīr-ē-ājam) and مسٹر (misṭar) that may have proper name next to it. With the help of Title Person and surname, the system is able to detect Person Named Entities easily. The system is also able to identify those person names which are not part of the gazetteer. We have collected 34 Title Persons and 102 Designations.

6. The NER system performs well when it can identify true Named Entities by resolving the ambiguities. Our NER System is able to identity true Person Named Entity based on various rules like if the system encounters any ambiguity in person name it will treat it as a special case and apply different rules to make sure that it is a true person name. For example when system encounter the word کمل (kamal) then system try to resolve the ambiguity of this word with help of postposition as surname or preposition where it may found Title Person or Designation. If there is no clue to identify as Named Entity then at last it check out the post position of ambiguity word likeکمل کو گھر پے کام تھا (kamal kō ghar pe kam tha.) the word (kō) کو give us clue that it may be the person name , So the system tag it as Person Named Entity(PNE).

7. Rule is also used to identify numbers that are non numerals like 'پانچ'چھے'چار' ' (Four, Five, and Six). The system able to tags three words as one Number Entity likeتین   پانچ   سو (Three Hundred Five).

8. Person name may have abbreviations in the place of first name of person like اے پی جے اے کلام (A P J A Klam). If the system is able to identify surname alone then it always try to find it as abbreviation name.

9. The system is able to find out and tag abbreviations like (C P U) سی پی زو (B B C) بی بی سی etc.

10. Organizations are tagged during gazetteer look up. We have insufficient data related to organizations so we have used some heuristics to identify ORG tags. For Example if text includes Org (vō pañjābī yūnīvrasiṭī kā ṭālī-ē-ilam hai.)وہ پنجابی   یونیورسٹی کا طالب علم ہے and we don't have this Organization in gazetteer then system apply rules to find and tag org as "Punjabi university".

7. Algorithm for Urdu NER

We have developed the system on Windows platform using Dot Net framework 4. System is using other available classes of framework to implement all features of our NER system. Like Tokenizer and Linked List class and its functions. We have developed many other modules for different rules used by the system. The system works in linear complexity.

1. Input text, through file upload or user may type text in given Text Field.
2. Normalization of Input Text
   1.1 Remove Extra spaces to single space.
   1.2 Remove special chars from end of the strings.
3. Gazetteer lookup
   3.1. Gazetteer lookup up for Locations, Terms, Brands, Abbreviations and Organizations Tags.
4. Tokenized and Normalized
   4.1. Tokenized the input Text word by word and search against the Gazetteer
5. Search for Date and Time tags
   5.1 Search for Number (numeral), Date, Time, Email and URL Tags.
6. Rule to tag Person Names
   6.1 Rule to detect Person Name with the help of Title Person Name, Designation and Surname without using Gazetteer.
7. Suffix stripping is used
   7.1 Suffix Striping Rules are used to Detect Location Names, Organizations, Izaafats, and Some types of Person names.
8. Find Person Names and Numbers
   8.1 Find more Person Names through Various rules and Gazetteer lookup.
   8.2 Rules are applied for names up to three words of length.
   8.3 Rule to detect Abbreviation Names.
   8.4 Rule is applied to resolve ambiguity in names.
   8.5 Rule is applied to find Numbers in non numerals form.
9. Rule is applied to find out Abbreviations which were not found during Gazetteer search.
10. Rules are applied to find out Organizations
    10.1 Those Organizations entity which were not found during Gazetteer lookup, rule will try to find out and tag them as organization entities.
11. Show tagged output to user along with untagged data.

Algorithm is self explained; still lightening some of its steps regarding Gazetteer look up. Gazetteer lookup is used to find out various Named Entities in text. We have collected Named Entities related to various fields i.e. Politics, Business etc. In algorithm's step 3 gazetteer look up is used for Locations, Terms, Brands, Abbreviations and Organizations Tags. All these tags are less ambiguous so the system tags them without applying any rule. System have gazetteer list related to these tags which are not ambiguous. In Step 4 system tokenized the input text and normalized the tokens for further processing. Step 6, 7 and 8 are used to tag Person Names. In Step 6, algorithm finds out person names based on Title Person, Designation and Surname. In this step system is able find out Person Named Entities without any gazetteer list. In Step 8, system used gazetteer list to find out person names and apply various rules to resolve their ambiguity. Some person names have patterns in its suffix or in prefix of the word. So, Step 6 of algorithm finds out this kind of person name by using suffix stripping.

8. Evaluation Metrics

Standard evaluation metrics for Information Retrieval includes Precisions, Recall and F-measure.

Recall: Relevant information extracted from text. Recall defined as:

Recall: = No. of correct answers given by system / Total No. of possible correct answers in text.

Precision: Actual correct answers returned by system. Precision defined as:

Precision: = No. of correct answer/No. of answers given

F-Measure: Balances of Recall and Precision by using a parameters **β**. The F-measure is defined as:

F-*measure*=$(\beta^2+1)RP/(\beta^2P+R)$

β is weighted as **β**=1. When **β**=1 F-measure is called F1-measure. The F1-measure is defined as:-

F1-measure=2*RP/P+R

9. Evaluation and Results

We have constructed two sets of test data. Test data is collected from different websites [19] of Urdu. Test data mainly include News from different fields like Politics, Sports, Business and Science. The reason of collecting News data is that because News data is always full of Named Entities. So, it gives challenging job to our NER system to identify all different kinds of NE tags accurately. Test data also include ambiguous data, our NER system tried to resolve ambiguities and tag only true entities. Mainly ambiguities are in person names and the system resolves them by applying different rules. For evaluating the system we have tested the system on two different test data sets. Both test data sets have news and articles related to different domain. Test data set 1 have data related to political news, some articles and short stories. Test data set 2 mainly have news related to science and business.

Test Case	Number of tokens	Domain
Test set 1	12032 tokens	News and articles related to politics.
Test set 2	150243 tokens	News data related to science topics, business news.

TABLE 2-Test Sets and along with number of tokens and their domain

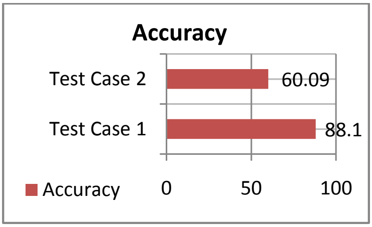

FIGURE 1- Performance of different test cases

Test Sets	Precision	Recall	F1-Measure
Test set 1	58.15	62.05	60.09
Test set 2	86.17	90.40	88.1

TABLE 3- Result of Test cases

Accuracy of Test set 1 is 88.1% and Test set 2 is 60.09%. Note that the frequency of occurrence of tags like Terms, Title Object, Brand Name are very less and we have not sufficient collected data related to these tags. Accuracy without all these four tags is shown below.

NE Tags	Precision	Recall	F1-Measure
NEP(Person Name)	92.85	93.37	93.10
NEL(Location)	85.00	90.00	87.28
NEO(Organizations)	77.70	81.30	80.37
NED(Designation)	86.80	89.21	87.98
NETP(Title Person)	85.33	88.45	86.85
NEM(Measure)	88.24	89.59	88.87
NEN(Number)	92.84	93.50	93.16
NETI(Time)	90.36	91.32	90.83
NEA(Abbreviation)	89.85	91.59	90.71

TABLE 4- Individual results of Nine NE tags

NE Tags	**Precision**	**Recall**	**F1-Measure**
NETE(Terms)	32.20	34.49	33.30
NEB(Brand Name)	43.45	47.28	45.34
NETO(Title Object)	51.49	53.22	52.35
NEIZ(Izaafats)	31.25	35.29	33.14

TABLE 5- Individual results of Four NE tags

Results shown in Table 5 for four tags are not good as compared to other tags because these tags need more time to collect accurate data. If we consider accuracy of all the thirteen tags then accuracy is 74.09%. Accuracy of the system also depends upon the domain of testing data. Test set 2 includes scientific and business terms so that system was not able to perform well. We have considered thirteen tag as compared to twelve tags used in IJCNLP-08. The system tagged Izaafat words because in Urdu, Izaafats are used very frequently and when we translate Urdu to target languages then we need to translate these Izaafats in specific target language words. Like (aab-e-hayat) آبِ حیات izaafat meaning in English is sacred water. But some time izaafat plays the role as NE for example وراثت خالصہ (vīrāsat-ē-khālsā) name of a place and it should not be translate in target language but transliterate it.

Conclusion and Future Work

We have developed the system to tag different Named Entities and system is able to find out and tag them all. But system has some limitations too.

1. We have developed the rule to tag person's name having length of three words. If person name has longer string of words like four words in a name then it will tag three words as one Named Entity and fourth one as another Named Entity. For Example in محمد آصف علی زرداری (mohmad āsīph alī jardārī) person name, زرداری (jardārī)will be tagged separately and محمد آصف علی (mohmad āsīph alī)will be tagged as another Named Entity. Count of Entities will be two in place of one.

2. Same is the case with Number tag. Like: دو سو چوّن (Two Hundred fifty four) tagged as one named entity but if we have longer string having more than three words like دو ارب چوّن لاکھ (Two Thousand Fifty Four Lakh) will be tagged as two separate entities. Where (Lakh) لاکھ will be tagged as a separately.

3. Partial tagging problem in Org tag like (انڈین انسٹیوٹ آف ٹیکنالوجی) Indian Institute of Technology will tagged partially as (Institute of Technology) word Indian will not be tagged with that Org tag.

4. Problem in Date tag, some time Date is written in word form like: بیس جنوری دویار گیارہ (Twenty January Two Thousand Eleven.) This kind of string will not be tagged as Date tag but will be treated as separate tags like January as Time tag and Twenty as Number tag and Two Thousand Eleven as one Number tag.

5. System is not using any kind of technique to resolve the segmentation problem in given Urdu text. Like other Asian languages Urdu has problem of space omission and space insertion. We will try to improve our NER system by using effective segmentation technique in near future.

6. System is not using POS Tagger or POS tagged data. So system always has to do larger number of comparisons to find out entities in given text. It is difficult to make any decisions for system based on rules because system compares only words but not their part of speech. Due to large numbers of comparisons system is little bit slowly when we gave large amount of input text to system. POS tagged data and stemmer is very essential for NER system and we have not used any POS tagger for our NER. So, we will develop POS tagger to include it in our system.

7. There is a problem to resolve ambiguity of person's name. We have collected the data of ambiguous person names and system treats them as special case to resolve its ambiguity but some time all the rules fail to remove its ambiguity. For example فتح کو بلاوا دو (phatah kō būlāva dō) here word فتح (phatah) can be proper noun or common noun. We do not have any clue to find out whether it is a person name or not. Rule checked Title Person, Designations or surname if all these conditions are not present in given sentence then we need to check out postpositions, but some time these postpositions come along with common name.

We have collected 6000 person names gazetteer. We will collect more number of person names and will try to include surname of other languages. Main problem is gazetteers of terms related to various domains. Collected data is not sufficient. We do not have standard spelling of terms related to various fields. We will try to collect standard spellings of different terms.

To develop the Urdu NER system we have insufficient resources as we discussed earlier but still we are able to get good accuracy. In future we will try to develop other essential tools for Urdu NER.

References

[1] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay.(2008). *"Language Independent Named Entity Recognition in Indian Languages"* In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing, pp. 33–40.

[2] Andrew Borthwick, John Sterling, Eugene Agichtein and Ralph Grishman *"Description of* the MENE Named Entity System as Used in MUC-7*"* http://acl.ldc.upenn.edu/muc7/M98-0018.pdf Accessed on December 2011.

[3] Amit Goyal.(2008) *"Named Entity Recognition for South Asian Languages"* In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, , Hyderabad, India. Asian Federation of Natural Language Processing, pp 89–96.

[4] Anil Kumar Singh.(2008) *"Named Entity Recognition for South and South East Asian Languages: Taking Stock"* In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages Hyderabad, India. Asian Federation of Natural Language Processing, pp 5–16.

[5] Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanyal.(2008). *"Named Entity Recognition for Indian Languages"* In Proceedings of the IJCNLP Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing, pp. 97–104.

[6] Anil Kumar Singh.(2008) "Named Entity Recognition for South and South East Asian Languages: Taking Stock" in Proceedings of the IJCNLP Workshop on NER for South and South East Asian Languages, pp 5–16.

[7] Ali Elsebai Farid Meziane and Fatma Zohra Belkredim.(2009). *"A Rule Based Persons Names Arabic Extraction System"* In Proceeding Communications of the IBIMA Volume 11.

[8] Bowen Sun *"Named entity recognition Evaluation of Existing Systems" Accessed* November 2011 http://daim.idi.ntnu.no/masteroppgave.pdf

[9] Duangmanee (Pew) Putthividhya.(2011). *"Bootstrapped Named Entity Recognition for Product Attribute Extraction"* In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. Association for Computational Linguistics, pp 1557–1567.

[10] David Nadeau, Satoshi Sekine. "A survey of Named Entity Recognition and classification" Accessed on November 2011 http://nlp.cs.nyu.edu/sekine/papers/li07.pdf

[11] Kashif Riaz.(2010). University of Minnesota Department of Computer Science Minneapolis, MN, USA *"Rule-based Named Entity Recognition in Urdu"* In Proceedings of the Named Entities Workshop, Uppsala, Sweden. Association for Computational Linguistics, pp 126–135.

[12] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma.(2008). *"Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition"* In Proceedings of the IJCNLP Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing, pp. 25–32.

[13] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma.(2008). "Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition" In proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing , pp. 52-32

[14] Praveen Kumar P.(2008). *"A Hybrid Named Entity Recognition System for South Asian Languages"* In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages,Hyderabad, India. Asian Federation of Natural Language Processing, pp.83–88.

[15] Pramod Kumar Gupta and Sunita Arora(2009) *"An Approach for Named Entity* Recognition System for H*indi":* An Experimental Study In Proceedings of ASCNT CDAC, Noida, India, pp. 103 – 108.

[16] Sujan Kumar Saha, Sanjay Chatterji ,and Sandipan Dandapat.(2008). *"A Hybrid Approach for Named Entity Recognition in Indian Languages"* In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India. Asian Federation of Natural Language Processing , pp. 17–24.

[17] Wenhui Liao and Sriharsha Veeramachaneni.(2009) *"A Simple Semi-*supervised Algorithm *For Named Entity Recognition"* In Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, Boulder, Colorado. Association for Computational Linguistics, pp. 58–65.

[18] http://en.wikipedia.org/wiki/Urdu Accessed on March 2012

[19] www.bbc.co.uk/urdu/ Accessed on March-May 2012