A Hybrid Approach for NER System for Scarce Resourced Language-URDU: Integrating n-gram with Rules and Gazetteers

SAEEDA NAZ*, ARIF IQBAL UMAR**, AND MUHAMMAD IMRAN RAZZAK***

RECEIVED ON 09.02.2015 ACEPTED ON 17.03.2015

ABSTRACT

We present a hybrid NER (Name Entity Recognition) system for Urdu script by integration of n-gram model (unigram and bigram), rules and gazetteers. We used prefix and suffix characters for rule construction instead of first name and last name lists or potential terms on the output list that is produced by n-gram model. Evaluation of the system is performed on two corpora, the IJCNLP NE (Named Entity) corpus and CRL NE corpus in Urdu text. The system achieved 92.65 and 87.6% using hybrid unigram and 92.47 and 86.83% using hybrid bigram on IJCNLP NE corpus and CRL NE corpus, respectively.

Key Words: Entity Recognition, Named Entities, N-Gram Model, Gazetteer Lists.

1. INTRODUCTION

NE is a proper name of anything present in e-text. According to linguistic study, proper name means a name of person, location, or organization etc. For examples "علامہ اقبال" Allama Iqbal is a person name, "ایبٹ آباد" (Abbottabad) is a location name, and "مسلم لیگ پارٹی" is an organization name. NER system takes natural language as an input and identifies all the proper names. Basically it is a two steps process:

(1) Identifying named entities or its boundary from the electrical text.

(2) Classifying them into predefined tagged categories e.g. Person names, Location names, Organization names, other categories and "none of the above".

Identification and classification of proper nouns in text is significant in several natural language processing applications such as Automatic Summarization, Machine Translation, Information Extraction, Information Retrieval, Question Answering, Text Mining and Genetics. Proper names are the target for information extraction and machine translation as they carry important information about the text itself. Performance of all NLP applications depends on NER system.

Identification of proper names is a tagging problem where the goal is to assign correct tag to each token. This identification and classification decides whether the lexical unit is the part of a proper noun phrase, if it is, then which category it belongs to. The "Named Entity" term was coined and promoted in the 6th and 7th MUC (Machine

* Department of Information Technology, Hazara University, Mansehra, KPK, Pakistan.
** Government Girls Post-Graduate College No.1 Abbottabad, Higher Education Department, KPK, Pakistan
*** King Abdullah University of Science & Technology, Riyadh, Saudi Arabia

Understanding Conferences) organized by DARPA (Defense Agency Research Project Agency). These conferences set the mile stone for NER system [1]. The concept of MUC-6 and MUC-7 was used in MET-2 (Multilingual Entity Task) for Japanese NER, CONLL (Conference on Computational Natural Language Learning) for Dutch and Spanish and CONLL for German. Arabic language also got attention in GALE (Global Autonomous Language Exploitation), which is a large scale project [2].

NER systems for English, European languages and some Asian languages (Chinese, Japanese etc.) have attained maturity level and are yielding accurate results [1]. These languages become rich resourced languages. Development of NER system for the SSEALS (South and South East Asian Languages) is a challenging task due to unavailability of resources and some features of language such as lack of capitalization and spelling variations [3]. Some work has been done on NER system for few Indian languages but very little computational research work has been initiated on NER for Urdu language. We have thoroughly examined the existing work on Urdu NER system and explained in detail the challenges that it is facing in [4]. The construction of accurate NER system for Urdu language is significant because Urdu documents are processed on Internet and it is spoken by a considerable population of the globe.

The language was emerged in 1100 AD. Literary the word "Urdu" means "horde" (Lashkar لشکر). This name was given to this language because it was spoken by warrior of Persian, Arabic, and Turkish origins who invaded India. It is national and official language of Pakistan and one of the official languages in five states of the India. It is also spoken in Bengal, UK, USA, UAE, and Canada. Its native speakers are 60.6 million and it is spoken by almost 490 millionpeople all over the world. It is ranked as 3rd largest language of the world [5]. Writing System of Urdu is Nasta'liq Script that is written from right to left with no capitalization like Semitic languages. Its characters are similar in shape to that of Arabic, Persian and Pashto language characters. Nature of Urdu language is highly agglutinative, inflectional and free words order.

The rest of the paper is organized as follows: Related work is presented in Section 2; in Section 3 challenges of Urdu NER are highlighted. In Section 4, Training and Testing Data are presented. In Section 5, proposed methodology is described. Sections 6, contains illustration of NER system. Results and comparisons are discussed in Sections 7 and Section 8 concludes the paper.

2. RELATED WORK

Three approaches namely rules based approach, statistical approach and hybrid approach are used for construction of NER system [3,6].

Rule based approach is also called Handcrafted Approach. It is based on seeking named entities in the text by using linguistic or handcrafted rules manually written by linguists. The rules require experience of grammatical knowledge of the language on part of the linguists. It also uses gazetteer lists. Rule based NER is preferred for language that has less resources and training data. Rule based NER system is capable of finding complex name entities which is not possible by statistical approaches. It provides an NER system that works well for specific language or domain with high accuracy but will not work for other language or domain. It is domain specific and is not transferable to other languages. It requires deep linguistic knowledge and its maintenance is also difficult.

Statistical approach is based on machine learning models like HMM (Hidden Markov Model) [6], ME (Maximum Entropy), ISL (Incremental Nonnegative Subspace Learning Scheme) [7], CRF (Conditional Random Field)[8], etc. It needs large amount of name entities tagged corpus

for training the machine learning models. All of the named entities and their types of training models should be labeled and they should be matched with the testing data on which the NER system will be run. Gazetteer lists and dictionaries are also used to classify words for achieving better results in the statistical approach. Statistical approach is not domain specific therefore; it is easily transferable and trainable to other languages or domains. Maintenance of ML based NER systems is also very easy and cheaper than hand based one.

Hybrid approach is based on the combination of the strongest points of rule based approach and statistical approach or combination of more than one model. The hybrid based NER systems are generally used for those languages which have rich morphology due to complex nature of languages. It yields result with high accuracy but it has a problem of transferring to other languages or domains due to linguistic rules.

To the best of our knowledge the first research paper on NER for digital Urdu text is [9]. It discussed some issues of Urdu language and developed Becker and Riaz corpus for Urdu language consisted of 2200 Urdu documents. NERSSEAL (Named Entity Recognition for South and South East Asian Languages) workshop of IJCNLP [10] held at IIT Hyderabad made efforts for development of NER systems for Bengali, Hindi, Oriya, Telugu and Urdu languages. Corpus of 36000 words were provided by [11] to all the researchers but the contribution of researchers for NER system for Urdu language could not materialized and no experiment has been reported.

Chatterji, et. al. [12] described NER system that used ME approach for Hindi, Bengali, Telugu, Oriya and Urdu. Linguistic rules and gazetteer lists were used to achieving better performance of NER for Hindi and Bengali languages. They achieved f-measures of 65.13, 65.96, 44.65, 18.74, and 35.47% in Hindi, Bengali, Oriya, Telugu and Urdu respectively. Gali, et. al. [13] presented conditional random field based NER for five languages namely Hindi, Bengali, Telugu, Oriya and Urdu. They aggregated machine learning approach with hand written rules or heuristics. The NER system is trained for Hindi and Telugu. The paper also introduced some linguistics issues related to SSEA languages. The system attained accuracy of 40.63, 50.06, 39.04, 40.94, and 43.46 f-values for Bengali, Hindi, Oriya, Telugu and Urdu respectively without sufficient resources.

Ekbal, et. al. [14] developed NER system using statistical CRF model for South and South East Asian languages, particularly for Bengali, Hindi, Telugu, Oriya and Urdu. Different contextual information and variety of features is used for finding and recognizing twelve classes of name entities in the system. The features were language independent for all the languages except Bengali and Hindi. They used rules for identifying nested NEs for all the five languages. They also used gazetteer lists for Bengali and Hindi. They obtained the F-measure of 59.39% for Bengali, 33.12% for Hindi, 28.71% for Oriya, 4.749% for Telugu and 35.52% for Urdu. Kumar and Kiran [15] presented NER system for five languages including Urdu using CRF, HMM and rules in IJCNLP workshop. The system obtained 39.77, 46.84, 45.84, 46.58, 44.73 f-measures for Bengali, Hindi, Oriya, Telugu and Urdu using rules with HMM and 35.71, 40.49, 36.76, 45.62 and 38.25% f-measures for Bengali, Hindi, Oriya, Telugu and Urdu using hybrid CRF model. Hybrid HMM model showed better performance than hybrid CRF model for all the languages.

Mukund and Srihari [16], and Mukund [17] developed information extraction system for Urdu language. They constructed a sub module of NER in information extraction system by using two models such as ME and CRF based NER for Urdu. The results of ME were 55.3% f-measures and CRF based module for NER showed improved f-measure value of 68.9%.

Riaz [18] presented rules based approach for NER in Urdu. The different rules formulated from 200 documents of Becker-Riaz Urdu corpus [19] and chose 600 documents out of 2,262 documents for better evaluation of the experiment. The system gave f-measure of 91.1% with 90.7% recall and 91.5% precision. He also used his rules based NER on 36000 Urdu words' corpus of NERSSEAL Workshop organized by IJCNLP in 2008 [10] and obtained f-measures of 72.4% without any changes in the sets of rule. The system improved the results of f-measures of 81.6% by developing new rules after observing the training set. The developed rule-based approach for NER in Urdu of this paper demonstrated encouraging result over all NER systems that used different approaches.

Recently, Singh, et. al. [20] contributed their efforts to develop rules based Urdu NER system using the IJCNLP corpus for thirteen NEs (twelve NE were proposed by [11] and one extra NEs).The overall accuracy is 74.09%. Another recent contribution [21] is the development of NER system using N-gram with gazetteer lists and showed 75.14% f-measure for N-gram and 75.83% f-measure for Bi-gram with list and back off smoothing technique using the ACL NE corpus.

3. CHALLENGES IN URDU NER

The construction of a robust Urdu NER is a complicated task because of the following limitations:

- No orthography capitalization of the initial letter of Urdu.
- Resources are scarce; there is no standard NE tagged Urdu corpus available.
- The Urdu language has agglutinative nature, to which some additional features can be added to the word to have more complex meaning, e.g. پشاور ➔ پشاوری (Peshawar to Peshawari).
- In Urdu Language, SOV (Subject Object Verb) word order may be used but usually the writers do not follow the word order.
- One NE may be written in various forms using different spellings in different situations even for native names.
- Some words are taken from other languages.
- The main issue is nested NE. The individual token may need more than one label for nested NE which makes the classification task difficult, e.g. محمد علی جناح یونیورسٹی (Muhammad Ali Jinnah University).
- A compound Named Entity is composed of multiple words. This brings more challenges to accurately detect the beginning and the ending of a multi-word NE, e.g. محمد علی جناح (Muhammad Ali Jinnah).
- Some entities are made up by using conjunction word such as اور (and), e.g. عمران اور عرفان کلنک (Imran and Irfan Clinic).
- A NE can be used as a person name or organization name or as a word other than nouns e.g. نور (Noor) is a name of person or organization and also equivalent to the English word "light".
- In Urdu it is quite difficult to recognize acronyms as NEs due to non-capitalization, e.g. بی بی سی (BBC) and یو۔این۔او (UNO).

4. PROPERTIES OF CORPORA

For testing and training purpose we have used two datasets IJCNLP NE and ACL NE.

IJCNLP NE Corpus: NERSSEAL workshop's corpus consists of total 48252 words (training words, 12805 testing

words) and 3611 NEs (2584 training NEs, 1027 testing NEs) [11]. This corpus was donated by CRULP to workshop organized by INJCNLP [10]. After modification and preprocessing, the NE tagged corpus size is reduced to 47650 tokens with 3238 NEs. The corpus is divided into training and testing data. The training dataset consists of 35339 tokens and 2312 NEs whereas the testing dataset comprises of 12311 tokens having 926 NEs.

ACL NE Corpus: An ACL NE tagged corpus consist of size 50936 [22]. In order to implement Urdu NER system, the corpus is reduced to 31279 tokens with 1526 NEs after NED tag correction. The corpus is divided into training and testing datasets. The training dataset contains 26363 tokens and 1304 NEs whereas testing dataset consists of 4916 tokens and 222 NEs.

Statistical models (Unigram and Bigram) are trained on the training data and later on gazetteers and rules are added in order to improve the accuracy. The specification of training and testing data of corpus-1 and corpus-2 NNEs is Non Named Entitlesare given in Tables 1-2.

As, IJCNLP NE corpus has been used for training the proposed system and annotated data are in the SSF (Shakti Standard Format) thus we have changed the SSF format data into <tag>XXX</tag> and we have tokenized the token on the white space base. IJCNLP NE corpus consists of 36K total number of words with more NEs for Urdu. There were several issues with the NE tagging that may affect training. In order to overcome this issue, we have modified the IJCNLP NE corpus.

TABLE. 1. A SPECIFICATION OF TRAINING CORPUS

Training Data	Total No. of Tokens	Total No. of NEs	Total No. of NNEs
IJCNLP NE Corpus	35339	2312	33027
ACL NE Corpus	26363	1304	25059

TABLE 2. A SPECIFICATION OF TESTING CORPUS

Training Data	Total No. of Tokens	Total No. of NEs	Total No. of NNEs
IJCNLP NE Corpus	12311	926	11385
ACL NE Corpus	4916	222	4694

5. HYBRID NER

We have used hybrid approach which applies unigram and bigram model, handcrafted rules and gazetteer list in order to identify and classify rigid names i.e. person, location, organization name, time and date. The model is trained with tagged NEs. The trained statistical model is used to train with tagged NEs and identify the NEs in the test data. Furthermore we have introduced handcrafted rules and gazetteer list in order to improve the results and solve the low recall problem. Proposed Hybrid Urdu NER system is depicted in Fig. 1 and the Algorithms 1-2 are illustrated.

5.1 Tokenization and Extraction of NEs

To train the unigram model the above training data is split and after splitting and applying regular expressions the NE in the above text will be in the following form.

5.1.1 N-Gram

We have used {word, </tag>} pair NEs to train the n-gram model. After tokenization, data is tested without tags using trained n-gram model. The n-gram models (unigram and bigram) assign the most probable NE tags to NEs and assign none to all other words based on probabilistic approach.

The Uni-gram model probability is described as in Equation (1):

$$P(w) = \frac{\text{count}(w)}{N} \quad (1)$$

Whereas count (w) is the number of times the word 'w' has been occurred in the training data and 'N' is the total number of words in the training data.

The probability calculated using bi-gram model is as in Equation (2):

$$P\left(t_i \middle| w_i, w_{i-1}\right) = \frac{\text{count}\left(w_{i-1}, w_i, t_i\right)}{\text{count}\left(w_{i-1}, w_i\right)} \quad (2)$$

Whereas count (w_{i-1}, w_i) is the number of times the bi-gram '$w_{i-1}w_i$' appears in the training data and count (w_{n-1}) represents the number of bi-grams starting with 'w_{n-1}'.

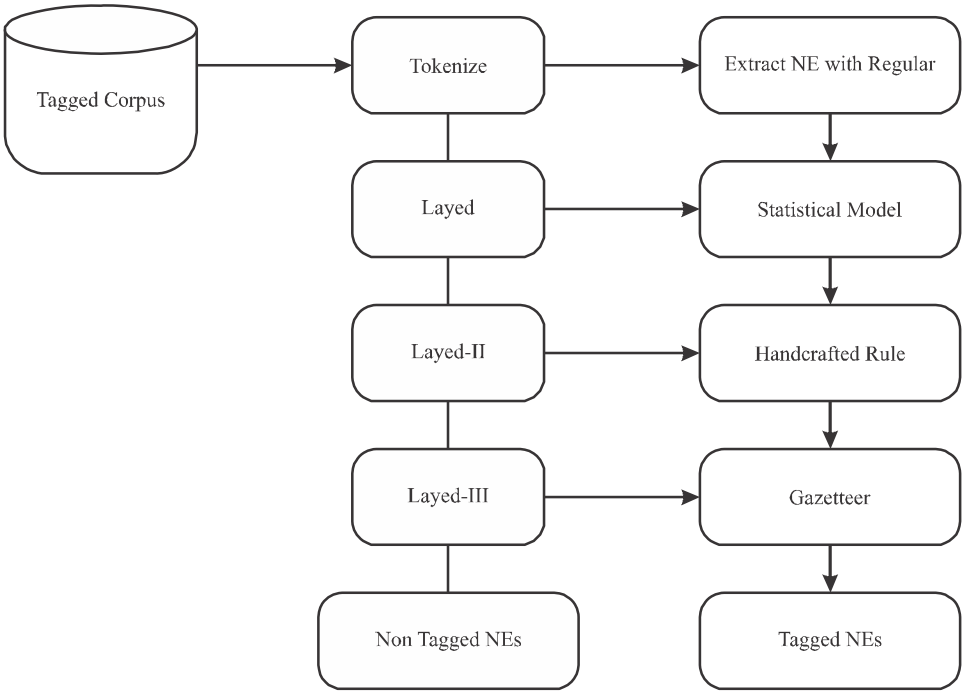


FIG. 1. ARCHITECTURE OF HYBRID URDU NER

ALGORITHM-1. HYBIRD UNIGRAM

UNIGRAM_NER( )
Input training_list, testing_list
Output result_list
Begin
Write training_list•**! INPUT_TRAIN( )**
Write testing_list•**! INPUT_TEST( )**
Compute P (t_i | w_i) from training_list fortraining the unigram model.
Test testing_list to trained unigram tagger for testing. //Unigram tagged the identified NEs with most probable NE tag and other with 'None'.
Write output_list//all words in the form 'word</tag>' result_list←
RULES_FUN(output_list)
Write result_list
End

ALGORITHM-2. HYBIRD BIGRAM

BIGRAM_NER( )
Input bigram_list, bitesting_list
Output result_list
Begin
Write bigram_list•!**INPUT_TRAIN( )**
Write bitesting_list•!**INPUT_TEST( )**
Compute P (w_it_i | w_{i-1}) from bigram_listfor training the bigram model.
Input bitesting_list totrained unigram tagger for testing. //Bigram tagged the identified NEs with most probable NE tag and other with 'None'.
Write output_list //all words in the form 'word</tag>'result_list←
RULESFUN(output_list)
Write result_list
End

5.1.2 Handcrafted Rules

Due to variations and the agglutinative nature of South East and Arabic script based languages, probabilistic graphical models result in less accuracy especially in case of recall [6]. To improve the accuracy and overcome the ambiguities, we have added the expert knowledge by introducing few handcrafted rules. The rules are developed according to the Urdu languages properties. The noon tags types are further analyzed and filtered with the help of rules. If any of the following rule trigger, then it append the appropriate tag with word.

R-I:

∀ " Words W_E, Extract W and append the location tag.m Where W_E ends with " ". For example:

الہ آباد </LOCATION> (Abbottabad-location name)

اسلام آباد </LOCATION> (Islamabad-location name)

R-II:

∀ " Words W_S, Extract W and append the date tag. Where W_S starts with " ". For example:

۱۵۸۳ ء سے </DATE> (Urdu Date format)

5.1.3 Gazetteer List

The gazetteer lists can be used to improve the recall. As compared to the other languages especially Latin script, work done for Urdu language processing is very less thus the language processing resources like gazetteers are not available. In order to add gazetteers, we have prepared several lists using different resources i.e. internet, training corpus etc. We have used gazetteers as a third filtration layer. If the word is not matched with rules and gazetteers then NNEtag is appended. Following lists are prepared for this work:

- List of person names (number of person NEs are 2526).
- List of location names (number of location NE are 1943).
- List of organization names (number of location NE are 506).
- List of date (number of location NE are 61).
- List of time (number of location NE are 44).

6. RESULTS AND DISCUSSION

Urdu NER task is performed using two statistical models; Unigram Model and Bigram Model. The results were not satisfactory due to rich language morphology and the issues of Urdu language that are discussed in section 3. To overcome these issues, we introduced several rules and gazetteers along with statistical models. We used multilayer approach (statistical method followed by rules and gazetteer respectively) in order to filter out tags.

We used the python's built-in Unigram tagger in the NLTK package. Training is performed on both corpora and then added rules and gazetteers according to Urdu language properties. We improved in precision of the NER system from 83.39-88.88%. Language based rules also increased the recall accuracy from 46.65-96.76%, this resulted in increasing the f-measure from 59.83-92.65% for IJCNLP NE corpus. But on ACL NE corpus, the precision is decreased from 90.00-78.36% whereas the recall accuracy and the f-measure increased from 40.54-99.54% that resulted in the increase of f-measure from 55.90-87.69% respectively.

The same is true for Bigram and hybrid bigram Urdu NER. A bigram NER tagger is implemented in python and training is performed on both corpora. With the addition of language based rules and gazetteers, the result was improved and hybrid Urdu NER has achieved 89.57, 95.57, and 92.47% precision, recall and f-measure respectively for corpus whereas we achieved 77.00, 99.54, and 86.83%

results for precision , recall and f-measure for ACL NE corpus, respectively. Table 3 summarizes results of proposed NER systems.

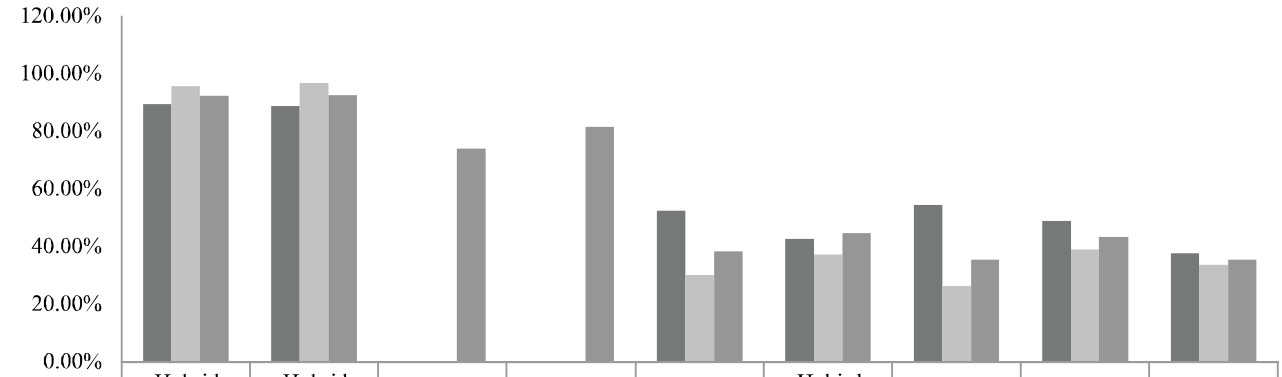
7. COMPARISON WITH EXISTING URDU NER SYSTEMS

7.1 Comparison of Existing Systems and Proposed NER System on IJCNLP NE Corpus

We compared the existing Urdu NER systems and proposed, Hybrid Unigram and Hybrid Bigram on IJCNLP NE corpus as shown in Fig. 2. Chatterji, et. al. [12], Gali and Surana [13], Ekbal, et. al. [14] and Kumar and Kiran [15] applied different statistical approaches and their results are not very promising. Riaz [18] improved statistical method by using rules based approach and got 81.6%. Fig. 2 shows, some of the promising results published in workshop of IJCNLP [10]. Chatterji, et. al. [12], obtained 35.47% f-measure accuracy for Urdu NER system using ME, whereas Gali, et. al. [13], Ekbal, et. al. [14] obtained 43.46, and 35.52% f-measure for Urdu NER system using CRF. Kumar and Kiran [15] used Hidden Markov Model and Conditional random Field and obtained 44.73, and 38.25% f-measure respectively on IJCNLP NE Corpus [11]. Fig. 2 shows that, accuracy can be improved by using language specific Gazetteers and rules. It is concluded that proposed methodology achieved considerable gain in accuracy as compared to the existing approaches from 81.60% f-measure to 92.65% f-measure and Hybrid Bigram. The "0" symbol in Fig. 2 means that overall precision and recalls were not given in the paper.

TABLE 3. A SUMMARY OF ALL RESULTS OF PROPOSED NER SYSTEM

Approaches	IJCNLP NE Corpus			ACL NE Corpus		
	Precision (%)	Recall (%)	F-Measure (%)	Precision (%)	Recall (%)	F-Measure (%)
Unigram	83.39	46.65	59.83	90.00	40.54	55.90
Hybrid Unigram	88.88	96.76	92.65	78.36	99.54	87.69
Bigram	98.39	19.87	33.06	92.30	16.21	27.58
Hybrid Bigram	89.57	95.57	92.47	77.00	99.54	86.83



	Hybrid Bigram (Proposed)	Hybrid Unigram (Proposed)	Rules based (Singh)	Rules based (Riaz)	Hybrid CRF (Kumar)	Hybird HMM (Kumar)	CRF (Ekbal)	GRF (Gali)	ME (Saha)
Precision	89.57%	88.88%	0	0	52.35%	42.81%	54.45%	48.96%	37.58%
Recall	95.57%	96.76%	0	0	30.15%	37.15%	26.36%	39.07%	33.58%
F-Measure	92.47%	92.65%	74.09%	81.60%	38.25%	44.73%	35.52%	43.46%	35.47%

FIG. 2. COMPARISONS BETWEEN OVERALL RESULTS OF EXISTING AND PROPOSED URDU NER SYSTEMS ON IJCNLP NE CORPUS

7.2 Comparison of Existing Systems and Proposed NER System on ACL NE Corpus

Similarly, the proposed hybrid Unigram and Bigram approach provided considerable f-measure gain of 92.65 and 92.47% as compared to 55.30% for ME and 68.90% for CRF respectively on ACL NE corpus. The detailed comparison on ACL NE Corpus is shown in Fig. 3. We use "0" symbol for the overall precision and accuracy of the system which were not given in the papers. Fig. 3 shows the gain in result as compare to the other approaches.

8. CONCLUSION

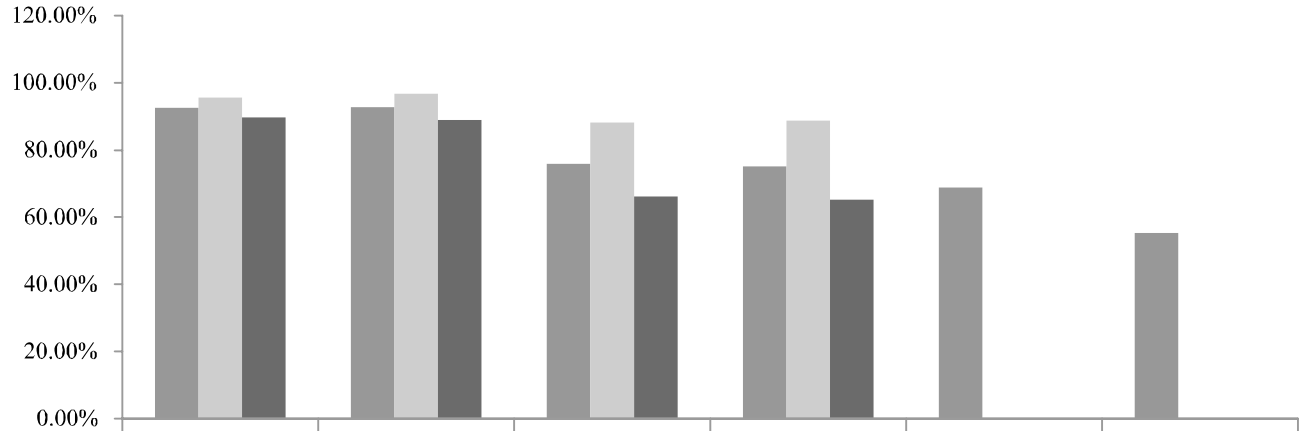
In this paper a hybrid NER system for Urdu script is presented by augmenting statistical model (unigram and bigram) with rules and gazetteer. For testing and training two dataset namely IJCNLP NE corpus and ACL NE corpus are used. The results show that presented approach attained considerable gain in accuracy when applied in layered manner. Additional language specific rule and gazetteer list helps to filter out the tags that cannot be filtered using statistical method due to the complexity of the script. The present work shows that addition of language rules along with statistical methods provided promising results for such morphologically rich languages, in future; we intend to improve the developed hybrid NER system by adding some additional rule and using CRF and ME model. We also intend to evaluate the proposed system on other scarce languages like Pashto, or Sindhi.

ACKNOWLEDGEMENTS

Authors are acknowledge and thanks to Mr. Waqas Anwar, for providing necessary information. Authors also highly appreciated to Mr. Sajjad Ahmed Khan, for his help in rules creation.

REFERENCES

[1] Nadeau, D., and Sekine, S., "A Survey of Named Entity Recognition and Classification", Lingvisticae Investig, Volume 30, No. 1, pp. 3-26, 2007.



	Hybrid Bigram (Proposed)	Hybrid Unigram (Proposed)	Bigram (Faryaal)	Unigram (Faryaal)	CRF (Mukund)	ME (Mukund)
F-Measure	92.47%	92.65%	75.83%	75.14%	68.90%	55.30%
Recall	95.57%	96.76%	88.18%	88.63%	0.00%	0.00%
Precision	89.57%	88.88%	66.20%	65.21%	0.00%	0.00%

FIG. 3. COMPARISONS BETWEEN OVERALL RESULTS OF EXISTING AND PROPOSED URDU NER SYSTEMS ON ACL NE CORPUS

[2] Huang, F., "Multilingual Named Entity Extraction and Translation from Text and Speech", Ph.D. Thesis, Carnegie Mellon University, Pittisburg, USA, 2005.

[3] Bouzoubaa, N., and Lachemi, M., "Self-Compacting Concrete Incorporating High Volumes of Class Fly Ash Preliminary Results", Cemement & Concrete Research, Volume 31, pp. 413–20, 2001.

[4] Naz, S., Umar, A.I., Shirazi, S.H., and Khan, S.A., "Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language", Research Journal of Applied Science Engineering & Technology, Volume 8, No. 10, pp. 1272–1278, 2014.

[5] "Ehtnologue: Statistical Summaries" (Last Visited: February, 2015).

[6] Bikel, D.M., Miller, S., Schwartz, R., and Weischedel, R., "Nymble: A High Performance Learning Name-Finder", Proceedings of 5th International Conference on Applied Natural Language Processing, pp. 194-201, 1997.

[7] Borthwick, A., "A Maximum Entropy Approach to Named Entity Recognition", Ph.D. Thesis, Departmemnt of Computer Science, New York University, USA, 1999.

[8] Li, W., and McCallum, A., "Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction", ACM Transactions on Asian Language Information Processing, Volume 2, No. 3, pp. 290–294, 2003.

[9] Becker, K.R., Bennett, B., Davis, E., and Panton, D., "Named Entity Recognition in Urdu: A Progress Report," Proceedings of International Conference on Internet Computing, 2002.

[10] http://mirror.aclweb.org/ijcnlp08/index.html.

[11] "Workshop on NER for South and South East Asian Languages, International Joint Conference Natural Language Processing, 2008. [Online]. Available: http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5" (visited: 2011).

[12] Chatterji, S.S., Dandapat, S., Sarkar, S., and Mitra, P., "A Hybrid Approach for Named Entity Recognition in Indian Languages", Proceedings of International Joint Conference on Natural Language Processing, pp. 17-24, Hyderabad, India, 2008.

[13] Gali, H., Surana, A., Vaidya, P., Shishtla, and Sharma, D.M., "Aggregating Machine Learning and Rule Based Heuristic for Named EntityRecognition", Proceedings of International Joint Conference on Natural Language Processing, pp. 25-32, Hyderabad, India, 2008.

[14] Ekbal, A., Haque, R., Das, A., Poka, V., and Bandyopadhyay, S., "Language Independent Named Entity Recognition in Indian Languages" , Proceedings of International Joint Conference on Natural Language Processing, pp. 33-40, Hyderabad, India, 2008.

[15] Kumar, P., and Kiran, R., "A Hybrid Named Entity Recognition System for South Asian Languages", Proceedings of International Joint Conference on Natural Language Processing, pp. 83-88, Hyderabad, India, 2008.

[16] Mukund, S., and Srihari, R.K., "NE Tagging for Urdu Based on Bootstrap POS Learning", Proceedings of 3rd Intenational Workshop on Cross Lingual Information, 2009.

[17] Mukund, S., "An Information-Extraction System for Urdu—A Resource-Poor Language", ACM Transactions on Asian Language Information Processing, Volume 9, No. 4, 2010.

[18] Riaz, K., "Rule-Based Named Entity Recognition in Urdu", Proceedings of ACL Named Entities Workshop, pp. 126-135, 2010.

[19] Becker, B., and Riaz, K., "A Study in Urdu Corpus Construction", Proceedings of 3rd Workshop on Asian Language Resources and International Standardization at the International Conference on Computational Linguistics, Taipei, Taiwan, 2002.

[20] Singh, U., Goyal, V., and Lehal, G.S., "Named Entity Recognition System for Urdu", Proceedings of International Conference on Computational Linguistics, Bombay, India, 2012.

[21] Jahangir, F., Anwar, W., Bajwa, U.I., and Wang, X., "N Gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language", Proceedings of International Conference on Computational Linguisics, Bombay, India, 2012

[22] "ACL NE Corpus" [Online]. Available: http://crl.nmsu.edu/Resources/lang_res/urdu.html (Visited: 2010-2011).