Named Entity Dataset for Urdu Named Entity Recognition Task

Wahab Khan[a]*, Ali Daud[b,a], Jamal A. Nasir[a], Tehmina Amjad[a]
[a]*Department of Computer Science and Software Engineering, IIU, Islamabad 44000, Pakistan*
[b]*Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
{wahab. phdcs72, ali. daud, jamal. nasir, tehminaamjad} @iiu.edu.pk

Abstract

Named entity recognition (NER) and classification is a very crucial task in Urdu. One challenge among the others which makes Urdu NER task complex is the non-availability of enough linguistic resources. The NER research for English and other Western languages has a long tradition and significant amount of work has been done to solve NER problems in these languages. From resource availability aspect Western languages are counted resource plentiful languages. On the other hand, Urdu lags far behind in terms of resources. In this paper we reported the development of NE tagged dataset for automated NER research in Urdu, especially with machine learning (ML) perspectives. The new developed Urdu NER dataset contains about 48000 words, comprising of 4621 named entities of seven named entity classes. The contents source of this new dataset is BBC Urdu and initially contains data from sport, national and international news domain. This new dataset can be used for training and testing purpose of various statistical and machine learning models such as e.g. hidden Markov model (HMM), maximum entropy (ME), Conditional random field (CRF), recurrent neural network (RNN) and so forth for conducting computational NER research in Urdu. Our goal is to make this dataset freely and widely acquirable, and to promote other researchers to exercise it as a criterial testbed for experimentations in Urdu NER research. In rest of the paper the new NER dataset will be referred as UNER dataset.

Key Words: Urdu, Named Entity Recognition (NER), Resources, Machine Learning (ML)

1. Introduction

Named Entity Recognition (also known as entity identification, entity chunking and entity extraction) is a task to identify and classify all proper nouns in texts into predefined categories, such as persons, locations, organizations, expressions of times, quantities, monetary values, etc. NER is an important subtask of many natural language processing tasks, such as information extraction, co-reference resolution, relation extraction, question answering, machine translation, etc [1-3]. NER system came in focus during the Message Understanding Conferences (MUCs) [3, 4]. After that plethora of NER techniques and systems are available [1, 5]. Most of these systems are developed for European languages [6], especially English and are fairly accurate. Many automatic frameworks for NER have been proposed for other non-European languages like Arabic, Persian and South Asian languages [7, 8]. For Urdu language, NER systems are yet in developing phase [9]. As most of the existing NER systems rely on the use of rich external linguistic resources e.g. annotated corpora, human-made dictionaries, gazetteers etc., to improve the system accuracy [10]. Urdu language lacks in having abundant linguistic resources. Characteristics of Urdu language make the NER task more difficult. For example, Capitalization is a bloom characteristic employed by NER systems for European languages Urdu language does not have Capitalization. Thus, Urdu NER task requires detailed analysis and adaptation.

2. The Urdu Language

In Pakistan more than sixty languages are verbalized, including a numeral of provincial languages. Urdu, also known as lingua franca, is the national language of Pakistan. Recently, Supreme Court of Pakistan also ordered the central Government of Pakistan to adopt Urdu as an official language throughout the country. Urdu is the main source of communication nationwide and is easily understood by about 75% people in Pakistan. But

statistics shows that Urdu is mother tongue of about 8% Pakistanis. Urdu is also a most popular language in India. In India Urdu is official language of six states and also the constitution of India recognizes Urdu as one among the 22 recognized languages. As per Wikipedia statistics, there are about 65 million native speakers of Urdu in both India and Pakistan. In Pakistan there are around eleven million Urdu speakers, 52 million native speakers are in India and world widely there are more than 300 million Urdu speakers [11, 12]. Pakistan, India, USA, U.K, Gulf countries, and Canada own Urdu speakers in copiousness. In the last few years NLP research community observed incredible fast growth of multilingual content on the web. As a result, NLP research community is attracted to explore monolingual and cross-lingual Information Retrieval (IR) tasks [3]. Initially, the web was designed to present information to users in English, but gradually with the passage of time, with the development of standard technologies and with opportunity of accessing the web resources uniformly around the globe, the web became multilingual source of information. Monolingual IR is centred on the queries and information present in same language, while cross-lingual IR is centred on the queries and information provided in numerous varied languages [13].

In the recent years, South Asian languages have gained intense research interest. Particularly, Urdu language is major research stake holder in Asian language processing research community [9].

NLP tasks such as part of speech (POS) tagging and named entity tagging etc. have paramount importance in all NLP systems [14]. NLP systems developed for English and other Western languages have criterial accuracies compared to Urdu [15-17].

3. Available Dataset

For automated Urdu language processing existence of bench mark datasets is mandatory. To train machine learning models, it is necessary to provide relevant pre-labeled training data and in large amount [3, 14]. As far as Urdu is concerned, it lacks in having abundant linguistic resources for development of Urdu NER tools and conducting experiments. The Center for Language Engineering (CLE)[8] in Pakistan has taken initiatives efforts in corpus-building activities as well in promoting and conducting research in Pakistani and many other Asian languages. For promoting research in Urdu, CLE has lunched many linguistic resources for Urdu language processing. Details of the available CLE linguistic resources can be found one the home page of CLE store http://www.cle.org.pk/clestore/index.htm. CLE provides all the linguistic resources with little amount of processing fee.

Although CLE provides a variety of linguistic resources but, so far, Urdu named entity tagged dataset is still missing from their linguistic resources list. CLE store have POS tagged dataset which is about 100K size. This POS tagged dataset can be used NER task to assign POS tags to words and then to use these POS tags as feature in training phase of ML models.

4. Related Work

The research work about Urdu NER task using machine learning techniques is still in initial stages. The core reason is the non-availability of standard NER linguistic resources. To accomplish supervised machine learning based Urdu NER task, a large pre-labeled NER dataset is mandatory [18], which is not available in case of Urdu. The research community from ULP domain is limited to use only two available NE tagged dataset for machine learning based research in Urdu NER task. The first one is the IJCNLP-2008 NE tagged dataset.

The IJCNLP-2008 dataset comprises of about 40000 words and in its annotation, twelve named entity classes are used.

The research group namely Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan and IIIT Hyderabad, India have jointly make efforts to create the IJCNLP-2008 dataset, after creation they donated it to the NER workshop [3, 12]. This dataset can be downloaded freely from its source UR: http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5. The second NE tagged dataset which we refer as Jahangir et al., [19], is a dataset of about 31860 words with total 1526 named entities. The dataset is annotated with four named entity classes. Details of these two dataset can be found in

[8] http://www.cle.org.pk/

Table 1. In Table 2 entity wise statistics are provided of Jahangir et al., and IJCNLP-2008 NE Tagged Dataset. The statistics are extracted from the available version of both the datasets with authors of this paper.

Table 1 Details of Two NE tagged datasets

Dataset	No. of Words	No. of Sentences	No. of NEs
Jahangir et al.,	31,860	1,315	1526
IJCNLP-2008	40408	1097	1115

Table 2 Entity wise Statistics

Entity	IJCNLP-2008	Jahangir et al.,
Person	277	380
Location	490	756
Organization	48	282
Date	123	101
Number	108	---
Designation	69	---

5. Development

Linguistic resources for most southeastern languages are not radially available due to which these languages are termed scared resource languages. Urdu is a southeastern language, and is spoken in a vast area of sub-continent. It is a low resource language. Due to resources scarceness not plenty of research work has been actioned for Urdu language [3].

The dataset we present contains all text from BBC Urdu cyber space. The current version UNER NE tagged dataset contains text from three news domain e.g. 1) national news 2) international news 3) sports news. In future, we will include text from other popular news domains e.g. entertainment, science, business, health not only form BBC Urdu but from other sources such as Express news, Dunia news etc. The size of our current NER dataset is about 0.48k words with total 4621 named entities. Entities in the text are manually tagged in the guide line of IJCNLP-2008 and Jahangir et al., dataset. Initially only seven named entity classes are used in tagging. The seven named entity classes used in tagging includes: PERSON, LOCATION, ORGANIZATION, DESIGNATION, NUMBER, DATE, TIME. After manual tagging the samples from the three domains are reviewed through Urdu linguistic experts from two different organizations, and changes mentioned by them are incorporated accordingly in the whole dataset. During the development process all the entities are tagged from right to left and also text is stored sentence level. The symbols "-" dash and "?" are used as sentence separators. During tagging entity are enclosed in start and end tags. E.g. The entity (پاکستان, Pakistan) where it occurs in text it is tagged with stat and end tags of LOCATION such as <LOCATION>پاکستان</LOCATION>. For all the seven entity class labels the same approach is adopted. For storage purpose we used notepad with UTF-8 encoding system. Text I n files are organized sentences wise because most of machine learning models takes inputs sentences wise. E.g. CRF, RNN, DRNN and so on.

The UNER NE tagged dataset we reported is freely available for research purposes and can be requested by sending an email to any of the authors. Although UNER NE tagged dataset can be used for rule based work, but its structure and organization is more feasible for machine learning based approaches. We hope that our NER dataset will help in promoting ML based research in Urdu particularly in NER task. Below Table 7,Table 8 and Table 9 shows single sentence from each news domain in which named entities are labeled with its corresponding class label.

Table 7 Structure of National News Domain

<LOCATION>پاکستان</LOCATION> کے صوبہ
<LOCATION>بلوچستان</LOCATION> کے دارالحکومت
<LOCATION>کوئٹہ</LOCATION> میں فائرنگ کے واقعے میں
<NUMBER>ایک</NUMBER> پولیس اہلکار سمیت
<NUMBER>تین</NUMBER> افراد ہلاک ہوگئے ہیں۔

Table 8 Structure of International News Domain

<LOCATION>پیرس</LOCATION> میں شدت پسند حملوں سے
عبدالقدیر حکیم <PERSON> منسلک ایک اور شدت پسند
قبل <TIME>دو روز</TIME> بھی <PERSON>مدنی</PERSON>
<LOCATION>عراق</LOCATION> کے شہر
<LOCATION>موصل</LOCATION> میں مارا گیا ہے۔

Table 9 Structure of Sports News Domain

<LOCATION>آسٹریلیا</LOCATION> کی جانب سے پہلی اننگز میں
<PERSON>سٹارک</PERSON> مچل
<NUMBER>5</NUMBER> وکٹوں کے ساتھ سب سے کامیاب بولر
ہے جبکہ <PERSON>ہیزل وڈ</PERSON> نے <NUMBER>

تین</NUMBER> اور سپنر <PERSON>لیون</PERSON> نے <NUMBER>دو</NUMBER> وکٹیں لیں۔

As most of the ML models require the training data in the format of IOB2 (Inside, Outside and Begin), IOE2 (Inside, Outside, and End) and SBME (Single, Begin, Middle and End) format [2]. The UNER dataset is fully compatible with the above mentioned format and one can easily convert into any format easily. Table provides an overview of SBME format of UNER dataset.

Table 6 Example of SBME format

Token	Type
پیرس	S_Location
عبدالقدیر	B_Person
حکیم	M_Person
مدنی	E_Person
دو	B_Time
روز	E_Time
عراق	S_Location
موصل	S_Location

While S means that this entity is single, B represents the beginning of an entity, M represents middle portion of an entity while the letter E represents the ending of an entity.

Details of our presented UNER dataset can be found in Table and

Table . Consolidated statistics such as total number of words, total number of comprising named entities and total number of sentences are provided in Table . Domain wise consolidated statistics of each named entity class are provided in

Table . Table provides lists of typical named entity types that are considered during construction process of our NE tagged dataset along with its description.

We intent to balance the dataset with consideration of genre, and proportion of each entity class**.**

Table clearly reflects that the proportion of DATE and TIME entity classes are quite small compared to others because the occurrence and mentioning these two entities in national, international and sorts news are not customary. After annotation process the whole dataset is stored in 150 notepad documents using UTF-8 encoding scheme. Table provides domain wise document detail of UNER dataset.

We believe that UNER dataset is a rich dataset and has ability to compensate the indigence's of NER and NLP research, utilizing the present-day Urdu.

Table 7 Consolidated Statistics of UNER Dataset

Total of No. of Words	48673
Total No. of sentences	1744
Total No. of Named Entities	4621

Table 8 Domain wise consolidated statistics of each entity class

Entity\Domain	Nat:	Inter:	Sport	Total
Person	401	201	605	1207
Location	390	360	455	1205
Organization	400	210	53	663
Designation	167	70	42	279
Number	270	132	589	991
Date	81	74	48	203
Time	40	23	10	73
Total	1749	1088	1809	4621

Table 9 Domain wise No. of Documents

Domain	File No.	No. of Document
National	1-60	60
Sports	61- 110	50
International	111- 150	40
Total		150

Table 10 List of Generic Urdu Named Entity Types with the kind of Entities they refer.

Type	**Tag**	**Sample Category**
Person	<PERSON>	Individuals, small groups
Location	<LOCATION>	Territory, land, kingdom, mountains, site, locality etc
Organization	<ORGANIZATION>	firms, group of players, Political parties, bureau etc
Designation	<DESIGNATION>	Various designations e.g. Professor, Dean, Mufti, Captain etc.
Number	<NUMBER>	Counts e.g.

		Hundred, Ten Thousand One, 10 million etc.
Date	<DATE>	Date stamps
Time	<TIME>	Clock time stamps

6. Conclusion

These days the state of the art approaches that are widely adopted around the globe for the development of NER tools in almost all languages, including Urdu are machine learning approaches. The core reason behind its wide usage is based on four features: a) the capability of automatic learning b) the degree of accuracy c) the speed of processing and d) generic nature. The basic need for ML approaches for training and testing is the availability of pre NE tagged dataset. As far as Urdu is concerned, it is termed as resource poor language. Therefore, in this work we tried to contribute in Urdu language resource with a large enough newly created NE tagged dataset. Significant efforts were made to build this huge NE tagged dataset compared to existing NE dataset with text from multi news domains. The fascination aspect of the UNER dataset is its size as well as its very rich NE contents. These two aspects make UNER dataset more feasible for ML techniques. We hope that this new dataset will spark light in ULP research community and will attract researcher in future to promote research in ULP.

7. References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, pp. 3-26, 2007.

[2] M. K. Malik and S. M. Sarwar, "urdu named entity recognition and classification system using conditional random field," *Science International* vol. 5, pp. 4473-4477, 2015 2015.

[3] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review,* pp. 1-33, 2016.

[4] B. M. Sundheim, "Overview of results of the MUC-6 evaluation," in *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996*, 1996, pp. 423-442.

[5] A. Roberts, R. J. Gaizauskas, M. Hepple, and Y. Guo, "Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation," in *LREC*, 2008.

[6] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 142-147.

[7] K. Shaalan and H. Raza, "NERA: Named entity recognition for Arabic," *Journal of the American Society for Information Science and Technology,* vol. 60, pp. 1652-1663, 2009.

[8] U. Singh, V. Goyal, and G. S. Lehal, "Named Entity Recognition System for Urdu," in *COLING*, 2012, pp. 2507-2518.

[9] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 9, pp. 1-43, 2010.

[10] J. i. Kazama and K. Torisawa, "Exploiting Wikipedia as external knowledge for named entity recognition," in *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 698-707.

[11] K. Riaz, "Baseline for Urdu IR evaluation," in *Proceedings of the 2nd ACM workshop on Improving non english web searching*, 2008, pp. 97-100.

[12] S. Hussain, "Resources for Urdu Language Processing," in *IJCNLP*, 2008, pp. 99-100.

[13] J. Capstick, A. K. Diagne, G. Erbach, H. Uszkoreit, A. Leisenberg, and M. Leisenberg, "A system for supporting cross-lingual information retrieval," *Information processing & management,* vol. 36, pp. 275-289, 2000.

[14] K. Riaz, "Rule-based named entity recognition in Urdu," in *the 2010 Named Entities Workshop*, 2010, pp. 126-135.

[15] F. Adeeba and S. Hussain, "Experiences in building the Urdu WordNet," *In proceesings of the 9th Workshop on Asian Language Resources collocated with IJCNLP, Chiang Mai, Thailand* pp. pp. 31-35, 2011.

[16] E. T. Al-Shammari, "Towards an Error-Free Stemming," in *IADIS European Conf. Data Mining*, 2008, pp. 160-163.

[17] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration,"

in *Conference on Language and Technology*, 2009.

[18] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait journal of Science,* vol. 43, pp. 66-84, 2016.

[19] F. Jahangir, W. Anwar, U. I. Bajwa, and X. Wang, "N-gram and gazetteer list based named entity recognition for urdu: A scarce resourced language," in *10th Workshop on Asian Language Resources*, 2012, pp. 95-104.