بسم الله الرحمن الرحيم

In the name of ALLAH, the Beneficent, the Merciful

Named Entity Dataset for Urdu NER Task

Presented by: WAHAB KHAN

Authors: Wahab Khan

Ali Daud

Jamal Abdul Nasir

Tehmina Amjad

Department of Computer Science and Software Engineering Faculty of Basic and Applied Sciences

International Islamic University, Islamabad

Contents

Introduction – Named Entity Recognition

- Named Entity Recognition is amongst the most basic of NLP tasks
- In literature it is referred with various names, e.g
  - **Entity identification**
  - **Entity chunking**
  - **Entity extraction**
- It corresponds to the identification and classification of all proper nouns in texts into pre-defined categories

NE as Classification Problem

- Assign a label "Y" to an Observation "X"

{X}

=

Person
Location
Organization
Date
Time
Number
Designation

{Y}

NE as Classification Problem

- Which NE Tag (Y) is this Word (X)?

X = لاہور

Person
Y=Location
Organization
Date
Time
Number
Designation

Input

Output

Introduction – Named Entity Recognition

The Beneficial – NLP tasks

1. Information Extraction
2. Question Answering
3. Machine Translation
4. Text Clustering
5. Co-reference Resolution
6. Relation Extraction

Challenges

- Named entity recognition (NER) and classification is a very crucial task in Urdu
- There may be number of reasons but the major one are below:
  - Non-availability of enough linguistic resources
  - Lack of Capitalization feature
  - Occurrence of Nested Entity
  - Complex Orthography

Motivations - Machine Learning

- The state of the art approaches adopted for development of NER tolls are based on Machine Learn (ML) models
- The core reason behind its wide usage is based on four features:
  - The capability of automatic learning
  - The degree of accuracy
  - The speed of processing and
  - Generic nature

Motivations - Machine Learning

- Large enough pre NE tagged dataset is pre-requisite for ML approaches
- ML based NER research for English and other Western languages has a long tradition
- From resource availability aspect Western languages are counted resource plentiful languages
- Urdu lags far behind in terms of resources when compared to Western Languages
- From ML perspectives Urdu NER is very less investigated

Objectives & Goals

Objectives

- In this paper we reported the development of NE tagged dataset for automated NER research in Urdu, especially with machine learning (ML) perspectives

Goals

- 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

Available Dataset

- The available dataset for ULP research community
  - The IJCNLP-2008 Dataset
  - Jahangir et al dataset
- IJCNLP-2008 dataset comprises of about 40000 words
- In Annotation twelve named entity classes are used
- Created after joint efforts made by:
  - Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan
  - IIT Hyderabad, India

IJCNLP-2008 dataset

- Jahangir et al is a dataset of about 31860 words
- Contains total 1526 named entities
- In annotation four named entity classes are used

Dataset	No. of Words	No. of Sentences	No. of NEs
Jahangir et al.,	31,860	1,315	1,526
IJCNLP-2008	40,408	1,097	1,115

Entity wise statistics

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

The UNER Dataset

- In this research paper we reported development of a new NER dataset which we refer as UNER dataset
- The UNER dataset contains all text from BBC Urdu cyber space
- Initially the UNER dataset contain text from three news domain
  - National News
  - International News
  - Sports news
- Size is about 0.48k words
- Contains total 4621 named entities
- Seven named entity classes are used in tagging

Tags Description

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

Development

- All tagging performed manually
- IJCNLP-2008 and Jahangir et al., datasets are used as guideline
- Tagged samples are reviewed through Urdu linguistic experts from two different organizations
- Text is stored at sentence level using
- For storage purpose we used notepad with UTF-8 encoding system
- Entities are enclosed in start and end tags such as <LOCATION>پاکستان</LOCATION>

Data Samples of UNER

- Data sample of National News Domain of UNER

<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۔

- Data sample of International News Domain of UNER

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

Consolidated Statistics of UNER

Consolidated Statistics of UNER dataset

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

Entity Wise Statistics of UNER

Entity Wise statistics of UNER dataset

Entity\Domain	National	International	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

Documents of UNER

Documents statistics of UNER dataset

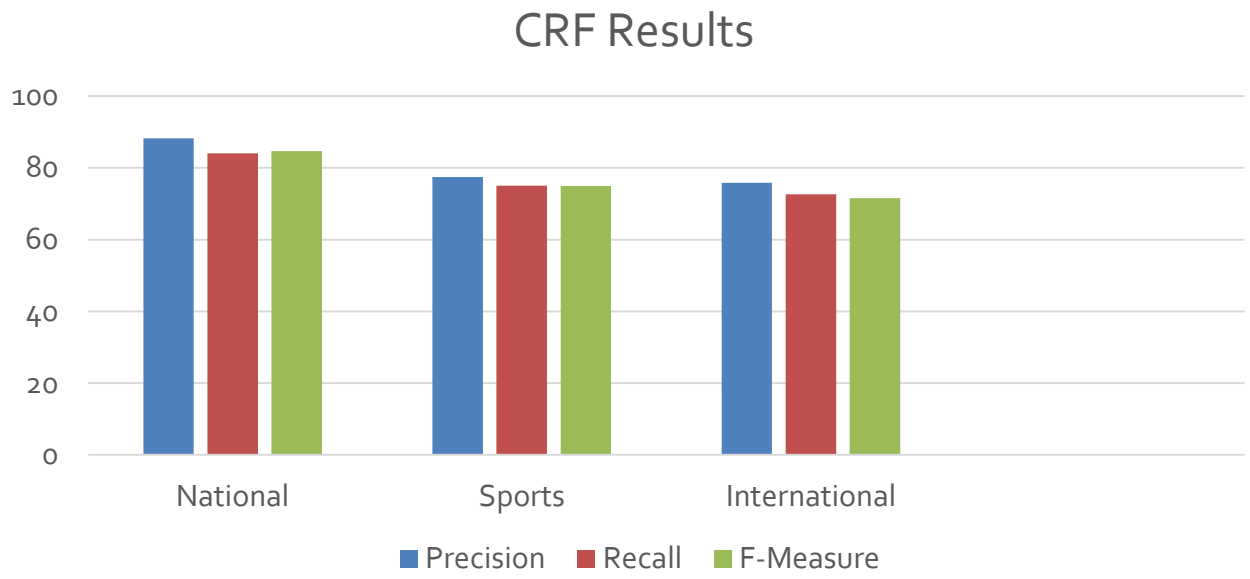
Domain	File No.	No. of Document
National	1-60	60
Sports	61- 110	50
International	111- 150	40
Total		150

Machine Learning models

- The UNER dataset can be used for training and testing purpose of various machine learning models such as e.g
  - Conditional Random fields(CRF)
  - Hidden Markov Model (HMM)
  - Support Vector Machine(SVM)
  - Recurrent Neural Network (RNN)

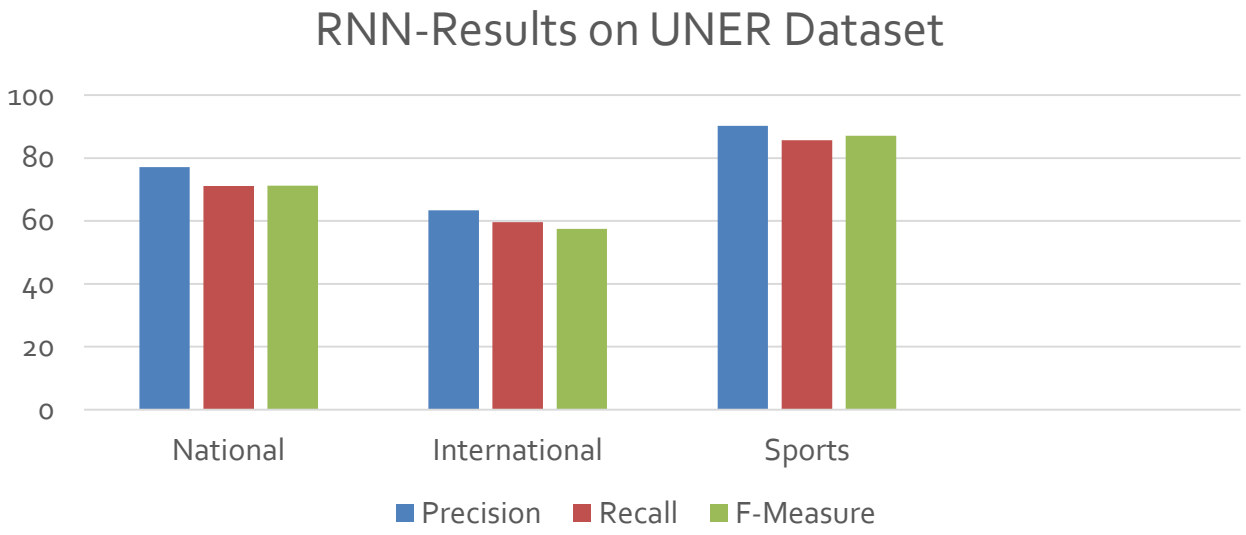
CRF-Results

Domain	Precision	Recall	F-Measure
National	88.21	84.05	84.68
Sports	77.44	75.02	74.92
International	75.84	72.62	71.56



RNN Results

Domain	Precision	Recall	F-Measure
National	77.14	71.11	71.26
International	63.42	59.59	57.45
Sports	90.23	85.65	87.09



Conclusion

- Urdu 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
- The two fascination aspect of the UNER dataset are:
  - Its size
  - Its very rich NE contents.
- We hope that this new dataset will spark light in ULP research community and will attract researcher in future to promote automated research in ULP.

Publications

Published:

- Daud, A., **Khan, W**. , and Che, D. 2016. Urdu language processing: a survey. *Artificial Intelligence Review*: 1-33. doi: 10.1007/s10462-016-9482-x (IF: 2.11)
- **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. (IF:0.30)

Under Review

1. Urdu Named Entity Recognition: A Deep Recurrent Neural Network Approach

   (Journal: Natural Language Engineering)

2. Urdu Named Entity Recognition: A CRF Approach

   (Journal: Language Resource and Evaluation)

3. Urdu Part of Speech (POS) Recognition: A CRF Approach

   (Journal: Quarterly Journal of Speech)

You are welcome ...

Questions ?

Comments !

Suggestions !!