7th INTERNATIONAL CONFERENCE ON
Language & Technology
19TH - 21ST FEB, 2020 www.cle.org.pk/clt20

Creating Urdu Named Entity Recognition (NER)

Description

By using the example of a NER system creation, the tutorial will explain how a machine learning based NLP application is created. The audience/participant will not only get an NER system at the end, but they will also be able to design similar systems.

NER is used for finding entities e.g, person, organization, date and location etc. from the text. It is the first step in the process of Information Extraction from the textual data.

We will pick an annotated dataset of Urdu Named Entities (NEs) and create different sets of features. We will use Part of Speech tags and Word Embedding as features, hence these concepts and their usage will be explained.

A machine learning system will be designed that will run experiments on different combinations of features. The result will be compared and merits/demerits of different kind of features will be discussed.

Aims and Learning Outcome

The participants/audience will learn the following:

- How is Urdu text processed?
- How is the text annotated?
- How are the features corresponding to annotated text created for machine learning?
- How is a machine learning model created?
- How is the best set of features and/or best model chosen?
- How Python (on Google Colab) is used to solve NLP problems?

Outline

1. NLP - overview of task and techniques 10 minutes
2. NLP concepts
    a. Named Entities 20 minutes
        i. Introduction

ii. Code Example - SpaCy
iii. Annotation - IOB Tagging

b. Part of Speech Tagging 10 minutes
   i. Introduction
   ii. Code Example - Urdu PoS Tagger(s)

c. Word Embedding 20 minutes
   i. Introduction
   ii. Code Example - Word2vec

3. Creating NER
   a. Creating Features from annotated dataset 20 minutes
   b. Applying Classification Algorithms 10 minutes
   c. Evaluation & Discussion 10 minutes

The programming will be performed in Python notebook (running on Google Colab). The dataset will be hosted on google drive.

Profile of the presenters

Dr. Tafseer Ahmed has teaching and research experience of more than 20 years in various institutes including University of Karachi, FAST NUCES Lahore, University of Konstanz, DHA Suffa University and Mohammad Ali Jinnah University. He is working on computational linguistics, text mining and machine learning. His goal is the development of solutions for Pakistani languages.

Dr. Tafseer Ahmed worked on computational grammar of Urdu as his postdoc work at University of Konstanz, Germany. He has worked on Urdu Propbank for University of Colorado, Boulder. He was co-PI of the DAAD, Germany funded project "Urdu Text to Speech".

https://www.jinnah.edu/member/dr-tafseer-ahmed/

Mr. M. Yaseen Khan received his academic degrees, BS and MS in Software Engineering from University of Karachi and PAF-KIET in 2010 and 2015 respectively. He had worked in software industry as well as in academia. As a software engineer, he has developed web applications, mobile applications and data science projects. Currently, he is a senior lecturer at Mohammad Ali Jinnah University, Karachi.

He is keenly interested in Text Summarization, Machine Learning, information retrieval, influence mining, affective computing, psychology, computational linguistics and deep learning.
https://www.linkedin.com/in/muhammadyaseenkhan/