Balochistan Journal of Engineering & Applied Sciences (BJEAS) – (p-ISSN: 2518-2706)
Volume 4, No. 1

Sindhi Parts of Speech Tagging Using Neural Networks

Adnan Ali[1], Abdul Samad[2], Sikandar khan[3], Muhammad Imran[4]
[1]Department of Software Engineering, BUITEMS, Quetta, Pakistan
[2]Dhanani Schools of Science & Engineering, Habib University, Karachi, Pakistan
[3]Department of Information Technology, BUITEMS, Quetta, Pakistan
[4]Department of Electrical Engineering, BUITEMS, Quetta, Pakistan
Corresponding Email: adnan.tahri@gmail.com

Abstract—Language is the method through which humans communicate, in either spoken or written form, and every language have some grammatical rules. Sindhi is one of the oldest languages that has some grammatical rules for parts of speech, and it is morphologically rich. Parts-of-Speech Tagging is a process of grammatically marking words in any natural language. In this research work, a deep learning approach on a Parts-of-Speech Tagger for Sindhi text is proposed. This tagger is built Long Short-Term Memory method (LSTM) and Gated Recurrent Unit (GRU) techniques. These types of approaches have never been used to tag Sindhi language. We used 79959 of Fast Text's pre-trained Sindhi word vectors with 300 dimensions. We used 7312 annotated corpora was developed from different sources (such as Sindhi books, stories, poetries and so on). It contains 459 sentences, 1584 distinct words. The data was divided into Training and Testing sets into 80% and 20%. The LSTM model outperformed the GRU's 94.50% by attaining 97.92% in training accuracy and 80.77% by attaining 85.80% in test accuracy. To the best of our knowledge, novelty of the presented work lies in the combinational approach for identification of POS tags in an enhanced version of the Sindhi corpus.

Keywords—Sindhi, Sindhi Language, Speech Tagging, Neural Networks