World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:12, 2021

ViraPart: A Text Refinement Framework for Automatic Speech Recognition and Natural Language Processing Tasks in Persian

Authors : Narges Farokhshad, Milad Molazadeh, Saman Jamalabbasi, Hamed Babaei Giglou, Saeed Bibak

Abstract : The Persian language is an inflectional subject-object-verb language. This fact makes Persian a more uncertain language. However, using techniques such as Zero-Width Non-Joiner (ZWNJ) recognition, punctuation restoration, and Persian Ezafe construction will lead us to a more understandable and precise language. In most of the works in Persian, these techniques are addressed individually. Despite that, we believe that for text refinement in Persian, all of these tasks are necessary. In this work, we proposed a ViraPart framework that uses embedded ParsBERT in its core for text clarifications. First, used the BERT variant for Persian followed by a classifier layer for classification procedures. Next, we combined models outputs to output cleartext. In the end, the proposed model for ZWNJ recognition, punctuation restoration, and Persian Ezafe construction performs the averaged F1 macro scores of 96.90%, 92.13%, and 98.50%, respectively. Experimental results show that our proposed approach is very effective in text refinement for the Persian language.

Keywords : Persian Ezafe, punctuation, ZWNJ, NLP, ParsBERT, transformers
Conference Title : ICCCS 2021 : International Conference on Computing, Communication and Security
Conference Location : Amsterdam, Netherlands
Conference Dates : December 02-03, 2021