低資源言語処理への機械学習および統計手法の適用に関する研究、事例：ダリ語とパシュート語

Form 3

Abstract

Title

A study on Applicability of Machine Learning and Statistical Approaches on Low Resources Language Processing: Case Dari and Pashto

低資源言語処理への機械学習および統計手法の適用に関する研究、事例：ダリ語とパシュート語

Research on processing regional and many Asian natural languages are state of the art in recent decade. Still, the study in the context of Dari and Pashto languages are imperceptible. This study concentrates on establishment of Dari and Pashto natural language processing systems. Besides, this study compares several novel and state of the art models to discover the most effective approaches. Moreover, this investigation developed eight corpora, 3 for Dari and 3 related to Pashto language, which is very useful for further research on Dari and Pashto NLP.

Nowadays, text classification for numerous purposes is becoming an essential task for relevant people. One of the main aims of this study is to establish a Pashto automatic text classification system in both single and multi-label classification contexts. To follow this, the author built a Pashto corpus which is a collection of Pashto documents due to the unavailability of publicly accessible related datasets.

In recent decades, automatic text summarization has become a state-of-the-art research area in natural language processing. This article is a novel work on automatic summaries of Dari and Pashto documents. It builds the first Dari and Pashto text corpuses that are a collection of online available articles. Experiments show that the proposed model is able to produce accurate summaries comparable to the relevant reference summaries.

Recently, extracting human beings' emotions for different purposes becomes a critical task for concerned businesses and organizations. This dissertation introduces novel models and compares their performance for discovering the polarity of Dari and Pashto texts by considering both positive and negative classes. To achieve this goal, we created a Pashto dataset of user-written texts due to the unavailability of any practical Pashto datasets for this study. Findings show that the proposed framework obtained the best performance of 98.5% accuracy in detecting sentiments from text.

In preceding decades, many researchers have focused on building and improving speech recognition systems to facilitate and enhance human-computer interaction. This work provides a more robust and less error-proven ASR system for Dari and Pashto through developing isolated words ASR using different novel deep learning techniques.

Name Dawodi Mursal