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

Abstract

Research on processing regional and many Asian natural languages are state of the art in recent decade. Particularly, the investigators emphasis on applying machine learning approaches in natural language processing area. Still, the study in the context of Dari and Pashto languages are imperceptible. Natural language processing of low resourced languages such as Dari and Pashto are very challenging. Dari and Pashto are the official languages of Afghanistan.

This work demonstrates Pashto and Dari natural languages computational linguistics. Consequently, it examines the various challenges and their implications in the processing of these languages. This study concentrates on discovering some computational linguistic solutions and establishment of natural language processing application systems in diverse areas for Dari and Pashto languages that make reasonable progress. It describes how to build effective corpora for low resourced languages and which preprocessing steps are useful to increase accurateness of algorithms for Dari and Pashto. Besides, this study compares several novel and state of the art models to determine the most effective approaches on low resourced languages such as Dari and Pashto.

Keywords: Dari, Pashto, Natural Language Processing, Low Resourced Languages, Computational Linguistics.