知の物理学研究センター / Institute for Physics of Intelligence

ipi seminar

Tarin Clanuwat

(Center for Open Data in the Humanities, National Institute of Informatics
/国立情報学研究所,人文学オープンデータ共同利用センター)

「Pre-modern Japanese Kuzushiji Character Recognition with Deep Learning」

２０１９年４月４日(木)　１０時３０分～１２時００分

東京大学本郷キャンパス理学部１号館９階　９１３号室

Kuzushiji, a cursive writing style, had been used in Japan for over a thousand years starting from the 8th century. Over 3millions books are preserved. However, following a change to the Japanese writing system in 1900, Kuzushiji has not been included in regular school curricula. Therefore, most Japanese natives nowadays cannot read books written or printed just 150 years ago. Museums and libraries have invested a great deal of effort into creating digital copies of these historical documents as a safeguard against fires, earthquakes and tsunami. The result has been datasets with hundreds of millions of photographs of historical documents which can only be read by a small number of specially trained experts. Thus there has been a great deal of interest in using Machine Learning to automatically recognize these historical texts and transcribe them into modern Japanese characters. In this presentation, I will talk about what is Kuzushiji and why do we need this research. Then I will also talk about the dataset, the recognition model and how we plan to make ancient documents more accessible to public users.