Recent Achievements of Deep Learning on Recognition of Modern Japanese Magazines

LE DUC ANH データ科学研究系 特任助教

Introduction

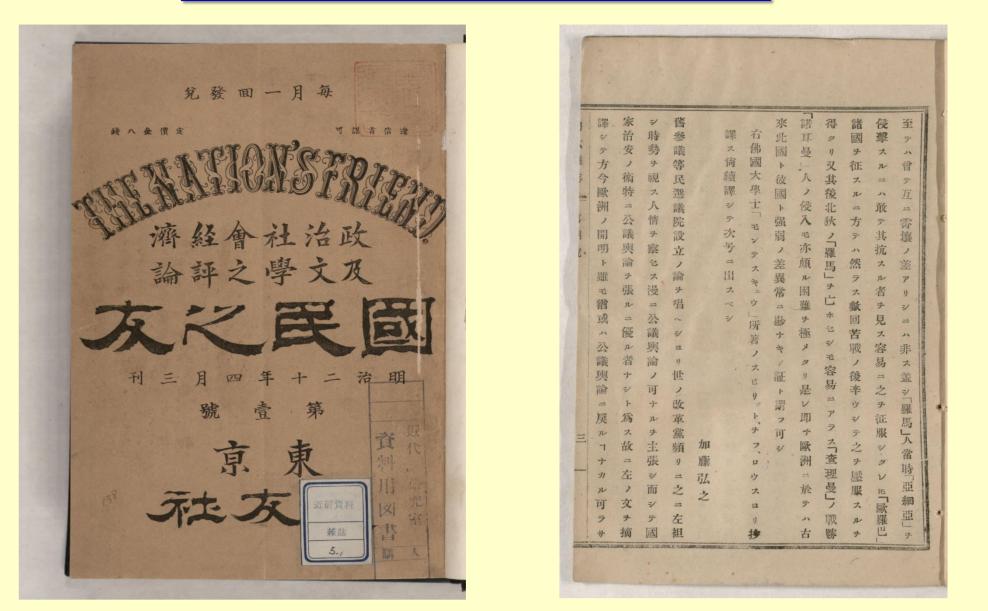
4 Since historical documents are an invaluable resource for historians in exploring social aspects, lifestyles, even weather in the previous era, many countries have been preserved their historical documents

4Document analysis and recognition can speed up the transcription process.

We Employ state-of-the-art technologies in Deep Learning to digitize modern Japanese document

We aim to build an OCR system for modern Japanese historical documents. The recognition system should be easy to train and reuse by other researchers from document analysis and digital humanities fields

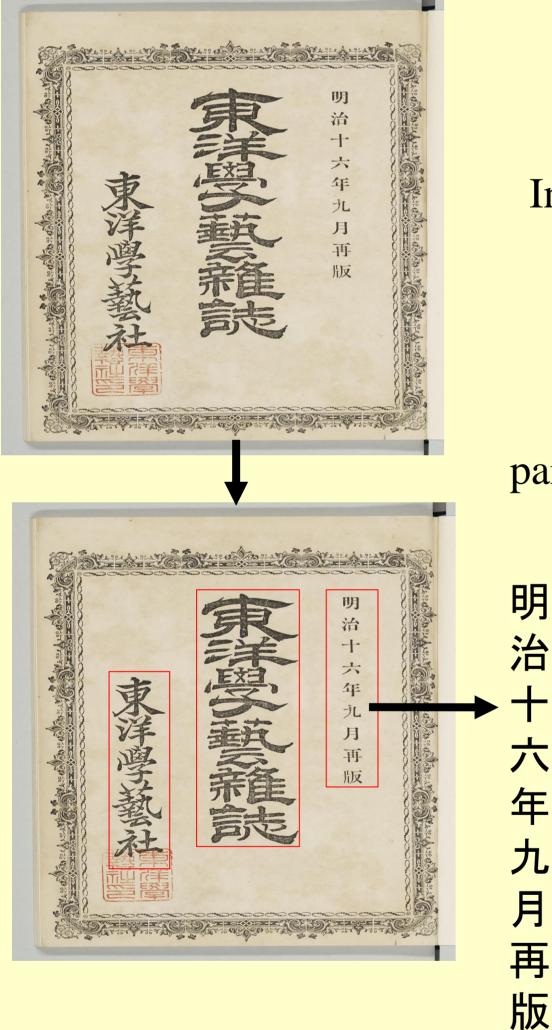
Dataset & results



Japanese Modern Magazines (~1900)

4Collected dataset: The dataset contains 922 pages from historical magazines in Japan from 1870 to 1945.

System Overview



Input image

Layout analysis paragraph, line, word

Text Recognition

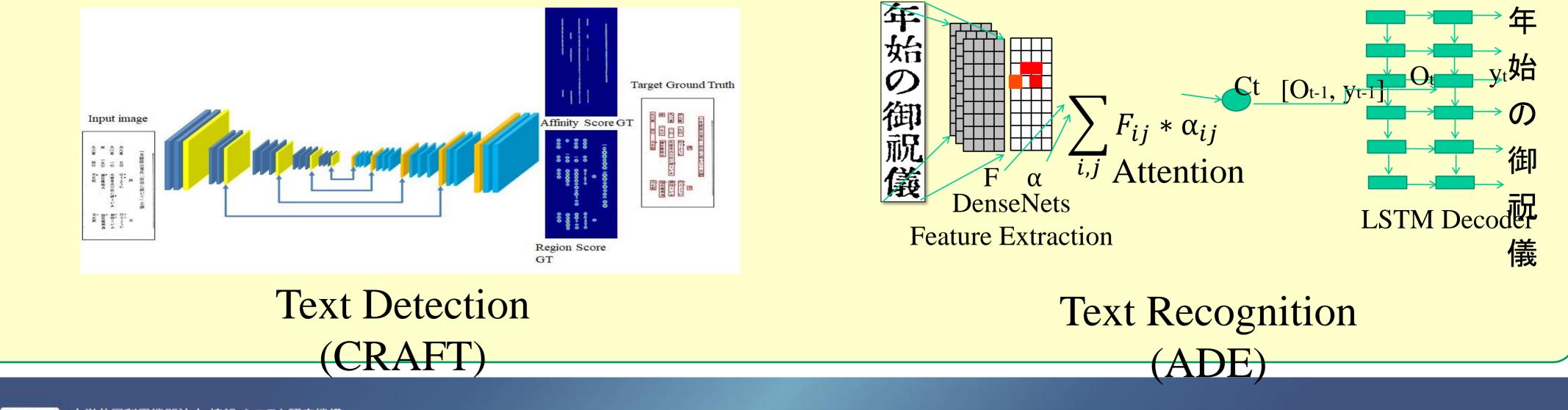
4The number of categories is 5,398 which contains many character categories that do not use in current Japanese character system.

]	Fext Dete	Text Recognition	
Precision	Recall	F1	Character Error Rate
89.6	92.0	90.8	22.52

Example of recognition result

_		r		-	0 0 0 0 0			
Ę	九三	號日	第友	之 民	國			
)解澤諭吉氏の演講筆記を読	典三動同盟の歐洲の やするに必要あるは し、なると見るの し、なると見るの し、今日	「あかかきいあるに係らず其の用意の局部にし其の意見或は少しく報知記者と施政の方針に關し其の意見の事が、報知記者と施政の方分に開し共のの見し報知新聞の外交政客し、報知新聞の外交政客見の目的何くにある	止らは復た是れ島田の振袖たるに過きさるならは彼た是れ島田の振袖たるに過ぎるならくの金環は月み類する如き貴族的の常用たらの金葉は月に類する如き貴族的の第用たらの金葉は月に類する如き貴族的の第用たらの金葉は月に類する如き貴族的の第用たらで、金属は月る如き貴族的の第一たの一個人見なし、	の如く細い向て	「「いたい」」「「いたい」」」」」」」」」」」」」」」」」」」」」」」」」」」	の思召書を		
https://github.com/ducanh841988/Kindai-0								

System Architecture



The Institute of Statistical Mathematics