An adaptive framework for end-to-end corrections in Indic-OCR

Optical character recognition (OCR) is the process of converting the document images into an editable electronic format. This has many advantages like data compression, enabling search or edit options in the images/text, and creating the database for other applications like machine translation, speech recognition, and enhancing dictionaries and language models. OCR in Indian Languages is quite challenging due to richness in inflections.

Using open source and commercial OCR systems, we have observed the word error rates (WER) of around 20-50% on typewriter printed documents according to our experiments. Also, developing a highly accurate OCR system with an accuracy as high as 90% is not useful unless aided by the mechanism to identify errors. So, we started with the problem of developing an end-to-end framework for error detection and corrections in Indic-OCR. We have outperformed state-of-the-art in 'error detection in Indic-OCR' for languages with varied inflections and have solved the out of vocabulary problem for 'error correction in Indic-OCR' in our ICDAR-2017 conference paper.

	100%		
amins 55 secs e	elapsed on this page(Right Clic	k to update)	
SLP1 Guide: # #	 अनन्तरापतविश्वविद्यालयः suggestions 		Raere
ब-a आग-A आग-I की-1 कि-1 कि-1 कि-1 कि-1 कि-1 कि-1 कि-1 कि	अनन्तशयनसंस्कृ	genos Cont+2 American Bedo Cont+5hift+2 American Cont Cont+X American	अनन्तरायनविश्वविद्यालयः
	ग्रन्थाङ्कः १८५.	Copy Link Location Paste Ctrl+V	अनस्तशयनसंस्कृतग्रन्थावस्तिः ।
	श्रीमटार्यभटाचार्या	Select All Ctrl+A	प्र≈याङ्कः १८५.
	आर्यभटीयं		श्रीमदार्थभटाचार्यविरचितम्
	गाग्यंकेरलनीलकण्ठसमिसुस्वविरचित- भाष्योपेतम्।		आर्यभटीयं
OCR Word		LSTM output/ Correct OOV Word	16000 Sanskrit: OCR vs LSTM 14000 LSTM 16000 Malayalam: OCR vs LSTM
एवमसक्तात्करणेऽवनिघ्नीः		एवमसकृत्करणेऽवनिसूनोः	12000
ല്ലാന്റെ	ടുത്തുകൊണ്ടിരിക്കും നടത്താണ്	ന്ന ക്ലാസ്സെടുത്തുകൊണ്ടിരിക്കു നട്ടർ പട്ടാ	5 10000
जसदपाऊ		जसदयाल	
g. 1. Exam ere, the con	ples of OCR words con rect words are all OOV	rected by LSTM in four Indic langua words. Mistakes are marked in red.	S. 2000 5 10 0 5 10 0 5 10 Edit Distance to Ground Truth
OCR Word		Correct Word	Percentage of Unique Word Occurences
ज्योतिःशास्त्रीवायकग्रन्थेषु ठिपका-कबडसा वगक्तिर		ज्योतिःशास्त्रीयविषयकग्रन्थे	
		ठिपका-कवडसा	S 80 Kanada Hindi
		वर्गाकार	on 60
Niruki«		Nirukta	j. 40
ig. 1. Exan bottom) v rrors are m	nples of OCR words in with no correct suggest arked in red and correct	Sanskrit, Marathi, Hindi and English tion from popular engines/spell-check tions in green. These errors are corre	$\frac{g_{0}}{g_{0}} = \frac{g_{0}}{g_{0}} = \frac{g_{0}}{g$

Prof. Ganesh Ramakrishnan, Department of Computer Science and Engineering, ganesh@cse.iitb.ac.in