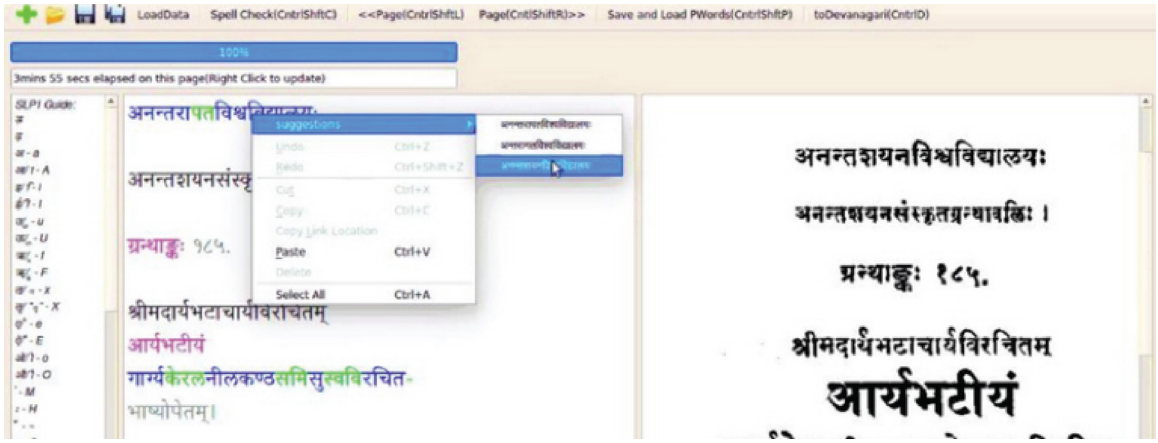
An adaptive framework for end-to-end corrections in Indic-OCR

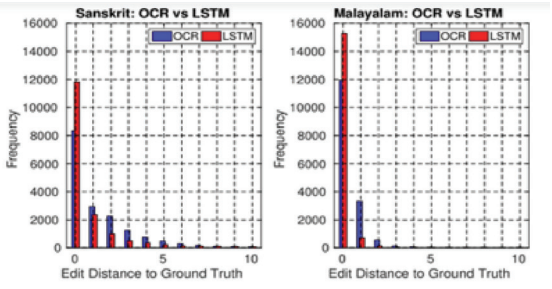
Optical character recognition (OCR) is the process of converting the document images into an editable electronic format. This has many advantages like data compression, enabling search or edit options in the images/text, and creating the database for other applications like machine translation, speech 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.



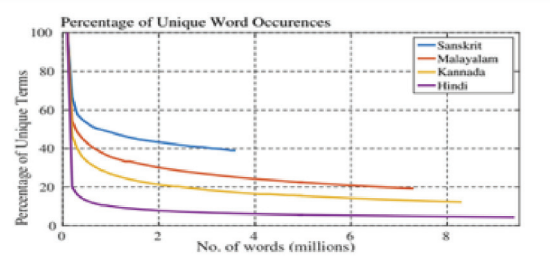
OCR Word	LSTM output/ Correct OOV Word
एवमसक्तात्करणेऽवनिघ्नीः	एवमसकृत्करणेऽवनिसूनोः
കല്ലാന്റെടുത്തുകൊണ്ടിരിക്കുന്ന	കല്ലാസ്സെടുത്തുകൊണ്ടിരിക്കുന്ന
ಗಜನಿವನ	ಗಜಸ್ಯವನ್ನು
जसदपाऊ	जसदयाल

Fig. 1. Examples of OCR words corrected by LSTM in four Indic languages. Here, the correct words are all OOV words. Mistakes are marked in red.



OCR Word	Correct Word
ज्योतिःशास्त्रीवायकग्रन्थेषु	ज्योतिःशास्त्रीयविषयकग्रन्थेषु
ठिपका-कबडसा	ठिपका-कवडसा
वगक्तिर	वर्गाकार
Niruki«	Nirukta

Fig. 1. Examples of OCR words in Sanskrit, Marathi, Hindi and English (top to bottom) with no correct suggestion from popular engines/spell-checkers. Errors are marked in red and corrections in green. These errors are corrected by our framework.



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