Supplementary material for - *Upcycle* Your OCR: Reusing OCRs for Post-OCR Text Correction in Romanised Sanskrit

Amrith Krishna[#], Bodhisattwa Prasad Majumder[*], Rajesh Shreedhar Bhat[**], and Pawan Goyal[#]

[#]Dept. of Computer Science and Engineering, IIT Kharagpur,
[*]Dept. of Computer Science, University of California, San Diego
[**]Walmart Labs, India
amrith@iitkgp.ac.in, bmajumde@eng.ucsd.edu,
rajeshbhatpesit@gmail.com, pawang@cse.iitkgp.ernet.in

1 Additional Training Details

We chose to train our own model for Romanised Sanskrit using the open source OCR Tesseract (Smith, 2007, 1987). Tesseract has been successfully used as an OCR for different languages and was an integral part of Google's digitisation efforts (Smith et al., 2009). Though tesseract does not provide a trained model for recognising Romanised Sanskrit, it provides a configuration file[1] which can be used for training a model. We attempted to train our own model using the configuration setting provided and used more than 80 different font styles for training.[2].

2 Missing Graphemes

Here is the complete list of 14 Sanskrit graphemes which are missing from the English vocabulary.

ā, ī, ū, ṝ, ḷ, ḹ, ṃ, ḥ, ṭ, ś, ṣ, ḍ, ṇ, ñ

3 Heatmap of mispredictions for different systems

Please refer to the ipython notebook for the detailed heatmaps with all mispredictions both OCR, PCRF and CopyNet. See ipython-notebook. URL: https://github.com/majumderb/sanskrit-ocr/tree/master/heatmaps

4 Synthetic Images

We have 7 settings of synthetic images, with which the training data has been built. All the 7 representative samples are given in Figure 1.

5 Results

The complete performance table for all the systems are presented in Table 1.

[1]https://github.com/tesseract-ocr/langdata/tree/master/iast

[2]Fonts obtained from http://www.pratyatosa.com/?P=41

6 Survey

We arranged a computerized survey to capture the human judgment for the system outputs. We sampled 15 unique ground truth test sentences from Sahasranāma. We obtained the OCR, PCRF and CopyNet system outputs of all the sentences. These 45 (15×3) strings are then arranged in 3 unique sets of questions, each having 15. Each set contains 5 strings from each of the three systems. None of two (or more) strings in a set belong to same ground truth. Apart from these 15 strings, each set also has one Ground Truth sentence, whose variant is not there in the set. This is to identify the normal response time of the expert for a correct sentence.

In the survey each page contains one predicted string and a question associated with it. Every page also contains a Yes/No question asking if the expert is familiar to the variant or not. For each set, we internally randomise the order of the pages for each survey instant. We capture the time taken per page as the expert takes the survey. Figure 2 contains the screen-shots of the welcome screen and pages containing survey questions.

Figure 1: Samples of synthetically generated images with all 7 distortion settings. The parameter settings (with SPN 0.5%; GN 2.5, refer Table 2 of original paper) are mentioned at the right side of the corresponding image. The bold-faced settings are the closest matches with the reference test datasets.

Model	Bhagavad Gīta			Sahasranāma			Combined		
	CRR	WRR	Norm LP	CRR	WRR	Norm LP	CRR	WRR	Norm LP
OCR	84.81%	64.40%	–	35.76%	0.65%	–	77.88%	23.84%	–
BiLSTM	93.79%	68.60%	-0.553	61.31%	7.28%	-1.292	85.23%	45.60%	-0.852
BiLSTM-CRF	94.68%	68.60%	-0.548	65.31%	7.28%	-1.281	85.82%	45.60%	-0.847
PCRF-seq2seq	**96.87%**	70.56%	-0.227	81.77%	9.34%	-1.216	87.94%	57.17%	-0.803
EncDec+Char	91.48%	68.00%	-0.542	63.63%	15.74%	-1.321	82.51%	47.37%	-0.865
EncDec+BPE	90.92%	68.00%	-0.496	61.53%	15.74%	-1.384	83.14%	45.98%	-0.842
CopyNet+BPE	96.90%	75.21%	-0.208	87.01%	31.34%	-1.121	89.53%	68.11%	-0.761
CopyNet+Alphabet	95.12%	74.23%	-0.267	85.31%	29.87%	-1.134	86.10%	66.09%	-0.798
CopyNet+Word	95.97%	73.09%	-0.198	85.62%	29.91%	-1.001	86.42%	66.54%	-0.610
CopyNet+BPE+Alphabet	**97.01%**	**75.21%**	**-0.165**	**87.01%**	**33.47%**	**-0.856**	**89.65%**	**68.71%**	**-0.551**

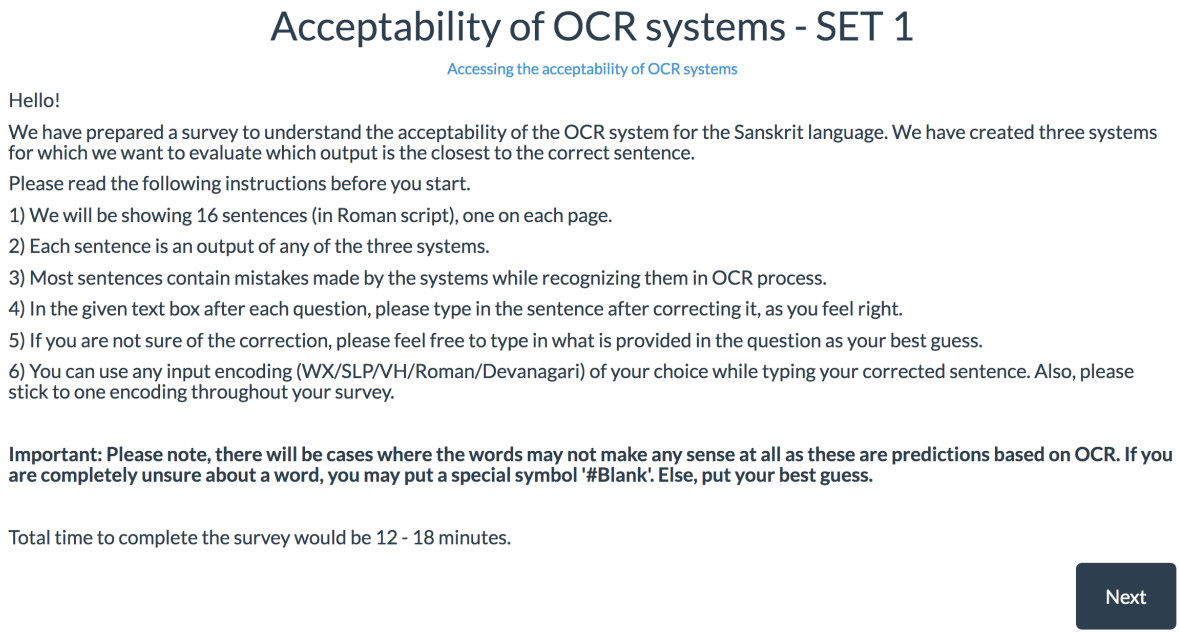
Table 1: Performance in terms of CRR, WRR and Norm LP (acceptability) for all the competing models

References

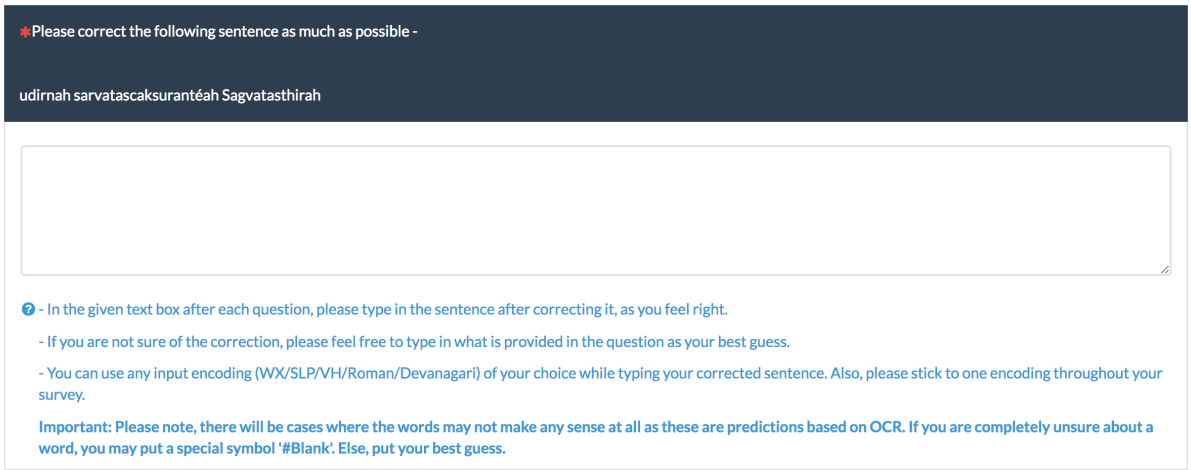
Ray Smith. 2007. An overview of the tesseract ocr engine. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 629–633. IEEE.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *Proceedings of the International Workshop on Multilingual OCR*, page 1. ACM.

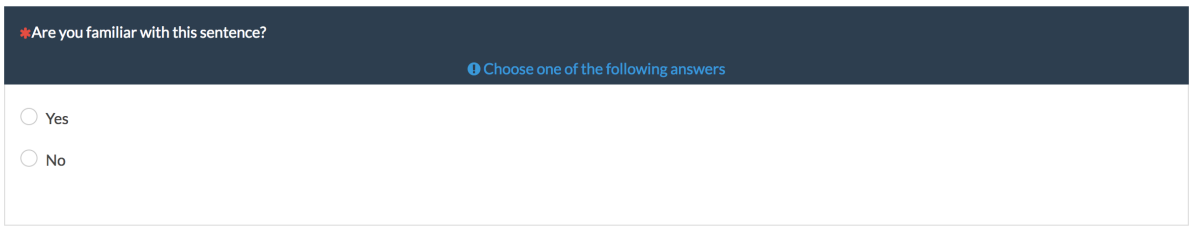
Raymond W Smith. 1987. *The Extraction and Recognition of Text from Multimedia Document Images*. Ph.D. thesis, University of Bristol.



(a)



(b)



(c)

Figure 2: Screenshots of the survey (a) Welcome screen (b) Page asking the question about one of the variants (c) the Yes/No question from the same page