MENTOR: Multilingual tExt detectioN TOward leaRning by analogy

Hsin-Ju Lin, Tsu-Chun Chung, Ching-Chun Hsiao, Pin-Yu Chen*, Wei-Chen Chiu, and Ching-Chun Huang

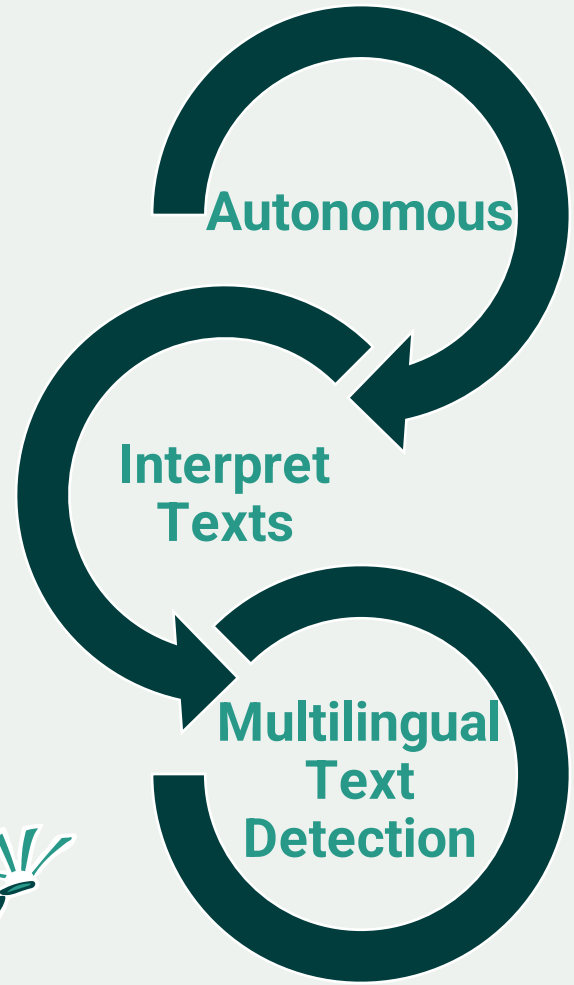
Organization: IBM research*
National Yang Ming Chiao Tung University

Reporter: Lucy Lin (Hsin-Ju Lin)

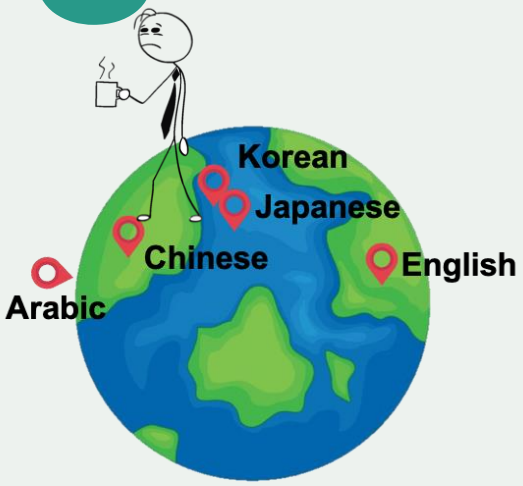
Outline

- Introduction
- Challenges
- Problem Setting
- Proposed Method
- Experimental Result
- Conclusion

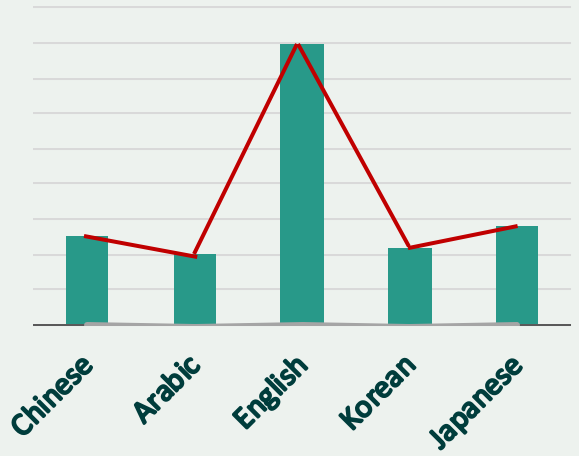
Introduction



3 Challenges

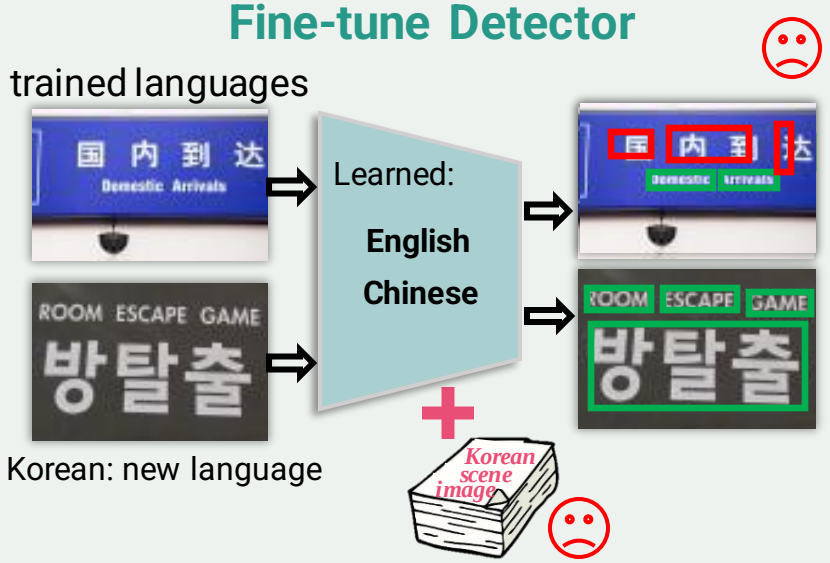


A universal text dataset containing all languages for supervised learning is not available.



Text training datasets that are imbalanced between languages.

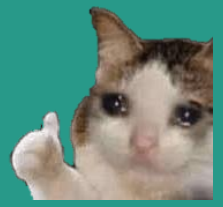
Fine-tune Detector



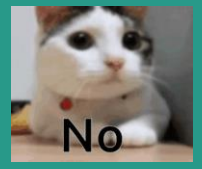
The detection model must be re-trained or fine-tune to detect new languages.

Problem Setting

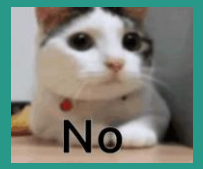
A generalizable multilingual text detector



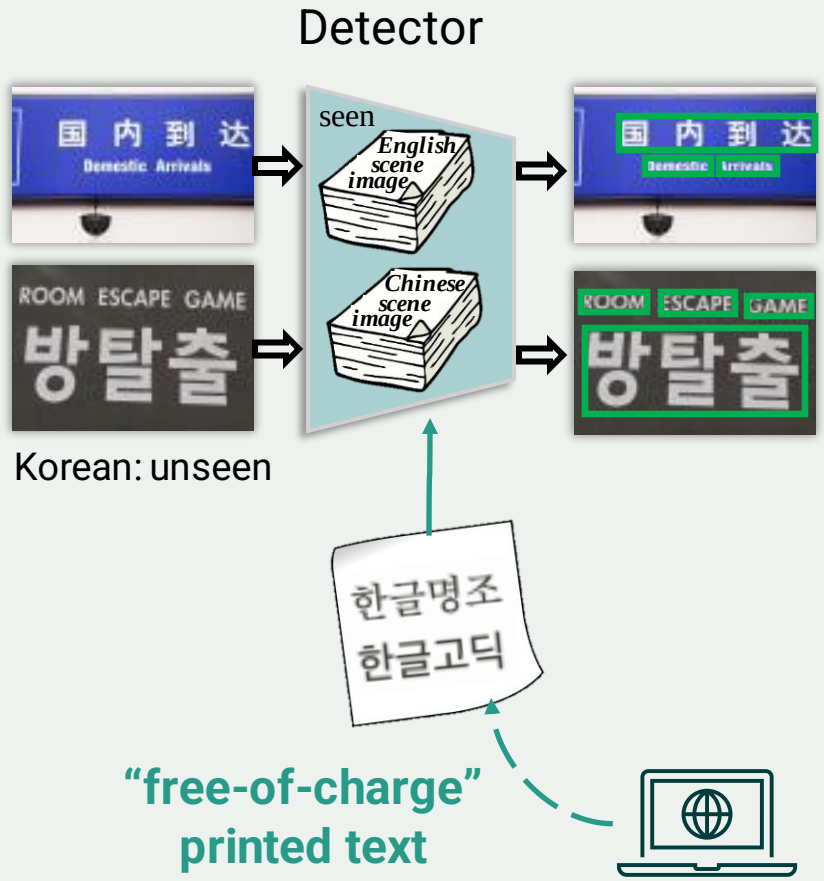
SHOULD detect both seen and unseen language regions



NO need to collect supervised training data for unseen language



NO model re-training required for unseen language



Proposed Method

printed text images generation

=

character set

가각갂갃간갅갆갇갈갉갊갋갌갍갎갏감갑값갓갔강
갖갗갘같갚갛개객갞갟갠갡갢갣갤갥갦갧갨갩갪갫
갬갭갮갯갰갱갲갳갴갵갶갷갸갹갺갻갼갽갾갿걀걁걂
걃걄걅걆걇걈걉걊걋걌갔걎걏강걒걓걔걕걖걗걘걙
걚걛걜걝걞걟걠걡걢걣걤걥걦걧걨걩걪걫걬걭걮걯
거걱걲걳건걵걶걷걸걹걺걻걼걽걾걿검겁겂것
겄겅겆겇겈겉겊겋게겍겎겏겐겑겒겓겔겕겖겗겘겙
겚겛겜겝겞겟겠겡겢겣겤겥겦겧겨격겪겫견겭겮겯
결겱겲겳겴겵겶겷겸겹겺겻겼경겾겿곀곁곂곃계곅
곆곇곈곉곊곋곌곍곎곏곐곑곒곓곔곕곖곗곘곙곚
곛곜고곡곢곣곤곥곦곧골곩곪곫곬곭곮곯……

character set + **font size** + **start position**

synthetic text generation

=

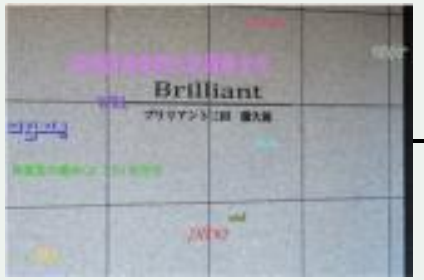
BBC news

BBC news + **font size** + **color** + **font style** + **position**

Proposed Method

Dynamic Guide

learn to generate representational attributes for different languages from printed text images



Text Finder

Obtain the differentiable features from scene images so that the text regions can be identified.

Language Mapper

learn the corresponding relationship between side information (printed text attributes) and the scene text feature.

Experimental Result

Quantitative evaluation (F1-score) on MLT17 dataset and Malayalam in IIIT-ILST dataset

Method	English	Arabic	Bangla	Chinese	Japanese	Korean	Malayalam
E2E-MLT [1]	55.43	55.431	3.027	50.594	12.9	32.715	x
MultiplexedOCR [2]	83.284	80.074	78.104	56.251	70.986	67.862	11.932
(a) Ours	84.031	80.952	81.76	83.585	76.896	72.479	65.837
(b) Ours	82.527	82.092	80.909	69.046	83.51	84.013	55.895
(c) Ours	81.778	54.646	82.134	82.554	80.737	83.143	43.992

[1] M.Bušta,Y.Patel,andJ.Matas,"E2e-mlt-anunconstrainedend-to-endmethodformulti- language scene text," ArXiv:1801.09919, 2018.

[2] M.Bušta,Y.Patel,andJ.Matas,"J. Huang, G. Pang, R. Kovvuri, M. Toh, K. J. Liang, P. Krishnan, X. Yin, and T. Hassner, "A multiplexed network for end-to-end, multilingual ocr," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, 2021.

Experimental Result

Quantitative evaluation (F1-score) on **synthetic** MLT17 dataset and **synthetic** Malayalam in IIIT-ILST dataset

Method	English	Arabic	Bangla	Chinese	Japanese	Korean	Malayalam
E2E-MLT [1]	50.19	54.67	4.027	55.138	22.346	34.317	x
MultiplexedOCR [2]	67.392	75.511	78.874	54.914	62.024	71.21	11.159
(a) Ours	71.76	74.949	79.044	74.351	71.556	50.534	55.607
(b) Ours	63.119	71.811	80.393	48.905	63.472	75.385	42.497
(c) Ours	63.424	46.061	80.328	80.328	59.254	75.385	36.586

[1] M.Bušta,Y.Patel,andJ.Matas,"E2e-mlt-anunconstrainedend-to-endmethodformulti- language scene text," ArXiv:1801.09919, 2018.

[2] M.Bušta,Y.Patel,andJ.Matas,"J. Huang, G. Pang, R. Kovvuri, M. Toh, K. J. Liang, P. Krishnan, X. Yin, and T. Hassner, "A multiplexed network for end-to-end, multilingual ocr," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, 2021.

Experimental Result

The visualization results of detecting language-agnostic text regions

- yellow bounding box means the text region belongs to unseen languages.

masking result

original image

Experimental Result

The detection results of synthetic MLT17 from different target languages

(a) Japanese

(b) Arabic

(c) Bangla

(d) Korean

(e) English

(f) Chinese (unseen)

Conclusion

1) We introduced a **new problem setting** for multilingual scene text detection and proposed a novel method “MENTOR” to **deal with text detection for both seen and unseen languages**.

2) Our **“Dynamic Guide”**, a dynamic and learnable module, and **“Text Finder”** module can **extract language specific features for seen and unseen languages** from printed text image and scene image, respectively.

3) We can identify the text regions of unseen languages via our **“LM”** module, by **comparing pixel-wise scene text features with** language-specific printed text feature, that is, **its mentor**.

4) Experiments show that our “MENTOR” can **achieve comparable results** with supervised methods **for seen languages** and **outperforms** other methods **in detecting unseen languages.**

Thank you for listening

Refer to the github for more detail

SCAN ME