An OCR Classifier for Republican Chinese Newspaper Text

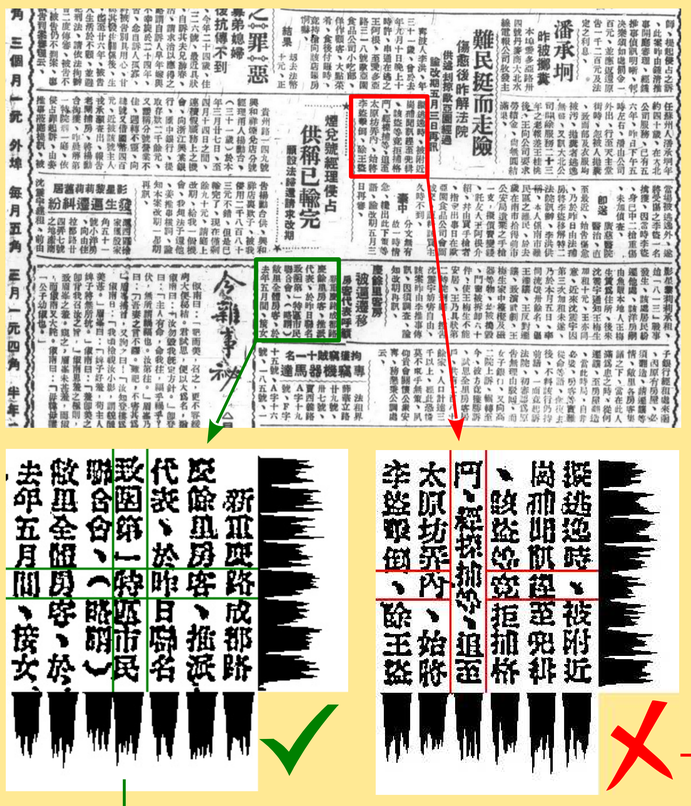
Konstantin Henke
Institute of Computational Linguistics

Matthias Arnold
Heidelberg Centre for Transcultural Studies

Early Chinese Periodicals Online

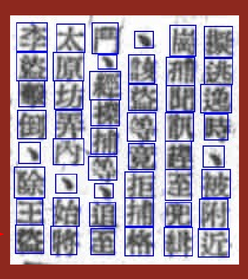
uni-heidelberg.de/ecpo

1 Manually cropping text blocks



We create a dataset of **text blocks without headings** by manually cropping them from the newspaper scans and assigning the **corresponding sections** of the **ground truth**.

We further create **projection profiles** from binaries to prepare the next step. Only text blocks with a coherent **grid-layout** are kept.



6 Outlook

Instead of grid-based character segmentation (cf. **2**), we will build on the **HRCenterNet** proposed by Tang et al. (2020).

segmentation result

新重慶路成都路
慶餘里房客、推派
代表、於昨日聯名
致函第一特區市民
聯合會、（略謂）
敝里全體房客、於
去年五月間、接女

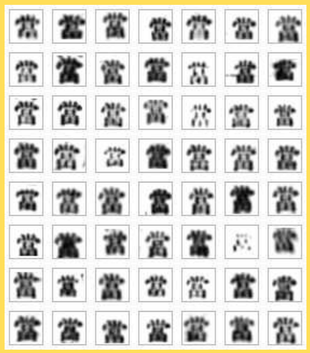
ground truth annotation

2 Character segmentation

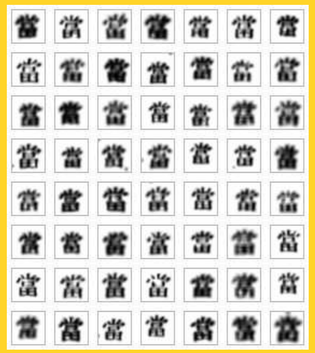
Building on the projection profiles, we cut the text blocks into **single character images**. These can easily be aligned with the ground truth (**1 field = 1 character**).

3 Character image generation

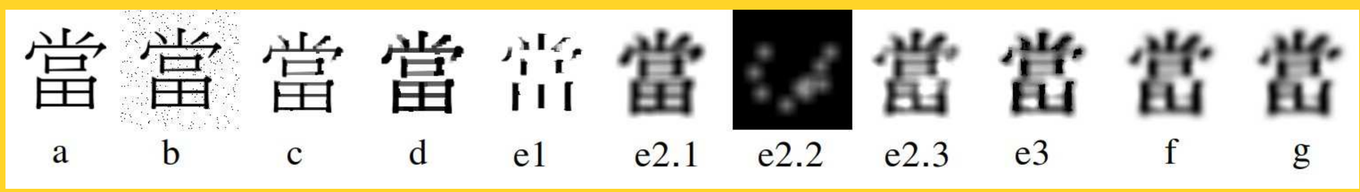
In order to **increase the size of the training set**, we use the method below to generate more character images from a **Song-Ti font**. All of the steps are **randomized augmentations** leading to a variety of new samples sufficiently **similar to the real images**:



samples of real images



samples of synthetic images



a Extract glyph images from font.
b Add random noise.
c Morphological opening and closing.
d Morphological erosion to thicken lines.
e 1 Extract vertical elements.
2 1 Separately erode and blur d.
2 Generate random patches.
3 Add the patches to the image.
3 Bitwise AND between e1 and e2.3.
f Blurring, random brightness change, rescaling to [0,255].
g Randomized elastic transformation.

4 OCR classifier (GoogLeNet)

We use the CNN architecture proposed by Szegedy et al. (2014) to train a classifier.

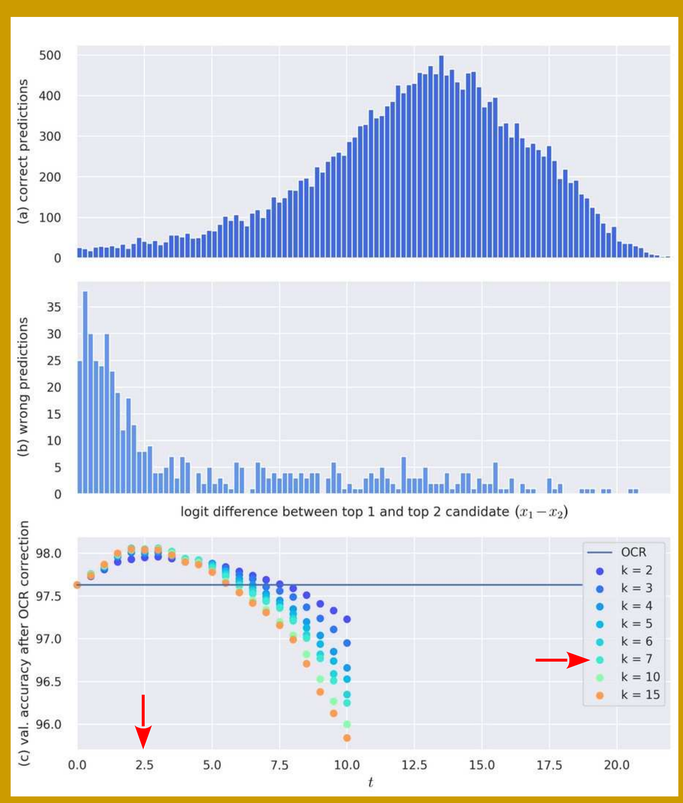
fine-tuning — pre-training

Results: (accuracy on test set) 95.49 % **96.95 %** 69.73 %

Remaining errors:

sample	Top 10 candidates output by the OCR classifier
骨	肯 貴 **骨** 片 督 昔 皆 貨 旨 嘗
貼	胎 **貼** 船 貼 晤 賠 始 販 斯 脂
汕	油 **汕** 別 遇 洲 勃 海 前 西 効
棒	棒 梭 慘 核 稼 橡 桂 俸 梓 控
薄	蒲 蓮 **薄** 通 滯 謝 浦 逝 蕩 鼎
歎	數 **歎** 歌 欸 歡 欽 教 歉 默 歔

5 Post-OCR error correction using a BERT model



Let x_1 and x_2 denote the **logit scores** of the **top 2 candidates** output by the OCR model. Set a **threshold** t. Any OCR prediction where $x_1 - x_2 < t$ (i.e. where the OCR model isn't "confident" enough its top candidate is correct) is passed on to a **pre-trained BERT model**. It uses the **given context** to **re-predict** the character choosing from the **top k OCR candidates**. Test for $t \in [0,0.5,\ldots,10]$ and $k \in [0,1,\ldots,18]$.

Result: (accuracy on test set) **97.44 %** for $t = 2.5$, $k = 7$

Heidelberg Centre for Transcultural Studies

Funded by:

Baden-Württemberg
MINISTRY OF SCIENCE, RESEARCH AND ARTS

Federal Ministry of Education and Research