An OCR Classifier for Republican Chinese Newspaper Text

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.

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:

- a) Extract glyph images from font.
- b) Add random noise.
- c) Morphological opening and closing.
- d) Morphological erosion to thicken lines.
- e1) Extract vertical elements.
- e2) Separately erode and blur d.
- f) Generate random patches.
- g) Blurring, random brightness change, rescaling to [0,255].
- h) Bitwise AND between e1 and e2.

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

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 repredict 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: 97.44% for $t = 2.5$, $k = 7$.

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

Konstantin Henke
Institute of Computational Linguistics

Matthias Arnold
Heidelberg Centre for Transcultural Studies

uni-heidelberg.de/ecpo