Deep learning-based NLP data pipeline for HER-scanned document information extraction

Jeong Hoon Choi
USCID: 5023184813

DSCI-560

Data Science Professional Practicum

Introduction & Related Work

- OCR is continuously evolving, incorporating methods such as Statistical Models (Machine Learning) and Deep Learning (i.e., LSTM).
- Traditional OCR involves the processes of segmentation, normalization, feature extraction, and classification.
- OCR technology and evaluation methods are less developed for the medical domain compared to other domains, and this paper studies OCR methods for EHR (Electronic Health Records)

OCR (Tesseract)
ML / DL) EHR (Electronic Health Records)

OCR : segmentation → normalization → feature extraction → classification (statistics / ML/DL

seperate background | minimize matrix size & reduce noise | feature vector

Methods

- Data source: UTMB HER
- Preprocessing: OpenCV
- OCR: Tesseract v 4.0.0
- Deidentification
- Segmentation
- Classification

Data Source: UTMB EHR

Image Preprocessing: pdf $\xrightarrow{\text{openCV}}$ gray-scale image

dilation/erode

Optical Character Recognition: Tesseract OCR v4.0.0

(extracted word + positional)

Deidentification: Pattern Match + Mask

Segmentation: Regular Expression (pattern match)

classification

- bag-of-word models: TF-IDF
- deep learning-based sequence models

Results

- Reports take many forms, including narratives in printed text, images, tables, and handwritings
- While it extracts data from text and tables well, its accuracy is relatively low for images and handwriting
- Overall accuracy is higher for deep learning-based sequence models than for bag-of-words models
- The ClinicalBERT model showed the best performance even with a small amount of training data

Question: What is the structure of the Deep Learning-based sequence model used in the paper?

Answer:

The architecture used in the paper has two main inputs.

It takes the words extracted from the previous OCR process and their positional information to create a sequence of appropriate token size, which is then fed into a sequence model.

Separately, structed data such as positional information and page numbers are fed into a feedforward neural network (FFNN).

Finally, the outputs of these two modules are processed through another FFNN to produce three labels (AHI, SaO2, Other).