Master IARFID
Reconocimiento de Escritura (RES)
Handwritten Text Recognition
Practical session:
from scanned pages to line images

Moisés Pastor & PRHLT-Group
mpastorg@prhlt.upv.es

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Description

- As the HTR engines can deal only with a line per image and the scanning is usualy done at whole page, it is mandatory to segment them.
- If the number of page is quite large, the manual segmentation becomes prohibitibe.
- We'll do in a semi-automatique way.

Rodrigo corpus

- Corresponds to a manuscript from 1545 entitled: "Historia de España del arçobispo Don Rodrigo"
- Written in old Castilian (Spanish) by a single author.
- It is a 853-page bound volume divided into 307 chapters describing chronicles from the Spanish history.
- According to experts, the manuscript writing style corresponds to Humanistic script, similar to the Italic script but with textual Gothic influences.
- Download it: `wget --no-check-certificate http://www.prhlt.upv.es/~mpastorg/RES/Rodrigo.tgz`
- Upacked it: `tar xvzf Rodrigo.tgz`

Tools

We'll need some software

- **imgtxtenh**: to clean the page images. From Mauricio Villegas git repository
  - `git clone https://github.com/mauvilsa/imgtxtenh`
  - `cd imgtxtenh`
  - `cmake .`
  - `make`
  - `cd ..`
- **baseLinePage toolkit**: to find where the lines are and extract them
  - `git clone https://github.com/moisesPastor/baseLinePage`
  - `cd baseLinePage/SRC`
  - `make`
  - `cd ../..`

Let's to start

- Set the path in order to find the executables:
  `export PATH=$PATH:$PWD/baseLinePage/BIN:$HOME/Pract/imgtxtenh`
- Create a dir for the clean version of the corpus: `mkdir data/Corpus_clean`
- Create the file list to be processed: `mkdir data/lists/; ls data/Corpus/*.jpg > data/lists/toClean.lst`
- Clean the images: `./neteja.sh data/lists/toClean.lst data/Corpus_clean` (2.5m. aprox.)

Training

- Label a page:
  - Copy the .xml from Corpus to Corpus_clean: `cp data/Corpus/*.xml data/Corpus_clean`
  - Go in the directory: `cd data/Corpus_clean`
  - Start the ground truth tool: `GT_Tool_PAGE_Points &`
  - Choose a xml page and label the baselines.
    1. Change to layout mode (F1) and label the text region,
    2. Change to baseline mode (F2) and label the baselines.
  - Get local minima points for this page:
    `imageLocalExtrema -i` *`ChangeForTheImageChoosedFileName`* `-w 15 -t 20 -k 2`
  - Load it into the GT_Tool_PAGE_Points
  - Tunne the settings and adjust the minima points
  - Go out the directory: `cd ../..`

Detection and Extraction

- Create a list file with the page to be used to train:
  `echo` *`FileNameWithoutExtensionNorPath`* `> data/lists/toTrain.lst`
- Train it:
  `baseLinePage/SCRIPTS/trainForestNPages.sh data/lists/toTrain.lst data/Corpus_clean 1 Rodrigo.cnf`
- Get the baselines (11m): `baseLinePage/SCRIPTS/getBaselines.sh data/lists/corpus.lst data/Corpus_clean data/Corpus_clean/Rodrigo_1.ert Rodrigo.cnf`
- Create the destination directory: `mkdir data/Corpus_clean_lines`
- Segment the images (18m): `baseLinePage/SCRIPTS/extractLines.sh data/lists/corpus.lst data/Corpus_clean data/Corpus_clean_lines &> err`

Results supervision and correction

Look for not well segmented pages, where the number of lines do not match with the transcripion one.

```
grep Region err| awk '{
  if (NF == 7) {
    N=split($1,NOM,"/");
    NLINS=$NF;
    cmd="wc -l data/txt/"NOM[N]".txt";
    cmd|getline sysOut;
    print NLINS,sysOut
  }
}' | awk '{
  DIF = ($1-$2);
  if (DIF != 0){
     if (DIF < 0)
        DIF=-DIF;
     print DIF,$NF
   }
}'|sort -nr|awk -F"/" '{print $NF}'|sed "s/.txt/.xml/">aRevisar.lst
```

Results supervision and correction

- Copy `aRevisar.lst` to `data/Corpus_clean`
- Go in the directory: `cd data/Corpus_clean`
- Correct them with: `GT_Tool_PAGE_Points -l aRevisar.lst`
- Go out the directory: `cd ../..`
- Remove the .xml extension: `sed -e "s/.xml//" aRevisar.lst > tmp.lst;`
- Remove the lines previously segmented:

```
for file in `cat tmp.lst`;
do
    echo \$file;
    rm data/Corpus\_clean\_lines/\${file}*;
done
```

- Segment the revised pages: `baseLinePage/SCRIPTS/extractLines.sh tmp.lst data/Corpus_clean data/Corpus_clean_lines &> err`