Master IARFID
Reconocimiento de Escritura (RES)
Practical session 4:
Indexing & Search on Handwritten Text Images

PRHLT-Group

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

KWS Approach

The KWS approach for indexing handwritten documents has the following features:

- Lexicon-based approach.
- Query set composed of single words.
- Line images are the basic spotting units; that is, we try to answer if a given word query is or is not in a line image.
- Word confidence score computation are based on the line level word graphs generated previously from the Rodrigo dataset.
- Performence evaluation based on *precision-recall* measures.

Most recent publication: `http://www.mdpi.com/2313-433X/4/1/15`

Required Tools

The following software will be used:

- **WordGraph2Index**: to produce probabilistic index from a word graph

```
mkdir -p Exp-KWS/WGIDX;
cd Exp-KWS

git clone https://github.com/PRHLT/WordGraph2Index.git
cd WordGraph2Index
make;
cd ..
```

- **kws-assessment**: to measure the spotting performance

```
git clone https://github.com/PRHLT/KwsEvalTool.git
cd KwsEvalTool
gcc -Wall -O4 -o kws-assessment kws-assessment.c;
cd ../..
```

Query Selection and Indexing

1. Selecting all keywords from transcripts with length greater than 1:

```
awk '{if (length($1) > 1 && $1 !~ /[<#]/ ) print $1}' \
   models/WFST/wordsMap.txt > Exp-KWS/keywords.lst
```

2. Obtaining index files from word graphs applying *posteriorgram* confidence score approaches.

```
for f in results/lattices/words/*.lat.gz; do
   F=$(basename $f .lat.gz); echo "Processing $F ..."; \

   Exp-KWS/WordGraph2Index/WordGraph2Index -i $f -z w | \
   awk '!/^#|NULL/{print $2,$3,$4,$5,$6}' > \
   Exp-KWS/WGIDX/${F}_p.idx; \
done
```

Building Index File for Evaluation

3. From each word in the word graph mark if appears on the line transcription. And generated entries of the form:

```
< lineId          word          [0|1]          [prob|-1] >
```

- 0 means the word on the word graph do not appear in the line ground truth.
- 1 means the word on the word graph appears in the line ground truth.
- -1 means that this word do not appears on the word graph but in the ground truth.

Building Index File for Evaluation

```
ls Exp-KWS/WGIDX/* > Exp-KWS/indx.lst

awk -v TR=data/text/test_words.ref ' BEGIN{
     while ((getline < TR) > 0)
       for (w=2; w<=NF; w++) REF[$1][$w]=1;
 }{
     lineId=$1;
     gsub("_p.idx", "",lineId)
     gsub(".*/", "",lineId)
     while ((getline word_line < $1)>0) {
         split(word_line, word);
         words_in_WG[word[1]]=1;
         result=0;
         if (word[1] in REF[lineId])
            result=1;
            print lineId,word[1],result,word[2]
     }
     for (w in REF[lineId]) #Words in REF but not in the WordGraph
         if (!(w in words_in_WG)){ print lineId,w,"1","-1"}
     delete words_in_WG;
}' Exp-KWS/indx.lst > Exp-KWS/IDX_p.dat
```

Performance Search Evaluation

4. Computing **AP** and **mAP**:

```
N_WG=`ls results/lattices/words/*.lat.gz | wc -l|cut -d ' ' -f 1`
# ---> 3322 lines

./Exp-KWS/KwsEvalTool/kws-assessment \
-t -s -a -m Exp-KWS/IDX_p.dat -w Exp-KWS/keywords.lst -l ${N_WG}
# MAP = 0.897 ( #Rel-Wrds = 2792 )
# AP = 0.919
```

5. Computing and drawing **R-P** curve:

```
./Exp-KWS/KwsEvalTool/kws-assessment -t -s Exp-KWS/IDX_p.dat \
-w Exp-KWS/keywords.lst -l 5011 > Exp-KWS/r-p_data.dat

cp ./Exp-KWS/KwsEvalTool/egs/plot-R-P.gnp Exp-KWS;
cd Exp-KWS/
gnuplot plot-R-P.gnp
evince R-P.pdf
```

Searching

4. Probabilistic Search

```
WORD_TO_SEARCH="castilla"
PROB=0.7

awk -v P=$PROB -v WORD=$WORD_TO_SEARCH \
  '{if ($2 == WORD && $NF > P) print }' Exp-KWS/IDX_p.dat
```

Evaluation