Search for Hyphenated Words in Probabilistic Indices: a Machine Learning Approach

José Andrés, Alejandro H. Toselli, Enrique Vidal

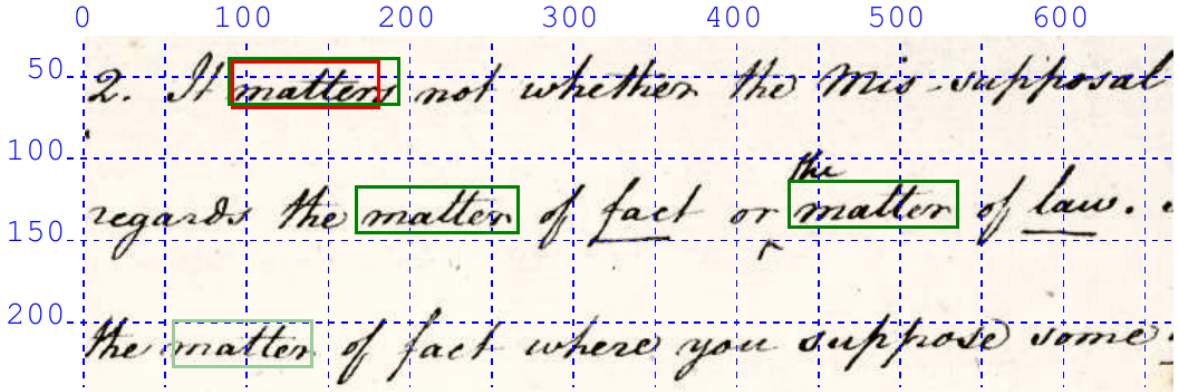
tranSkriptorium AI, Valencia, Spain

August 21st, 2023

tranSkriptorium

Introduction

Introduction



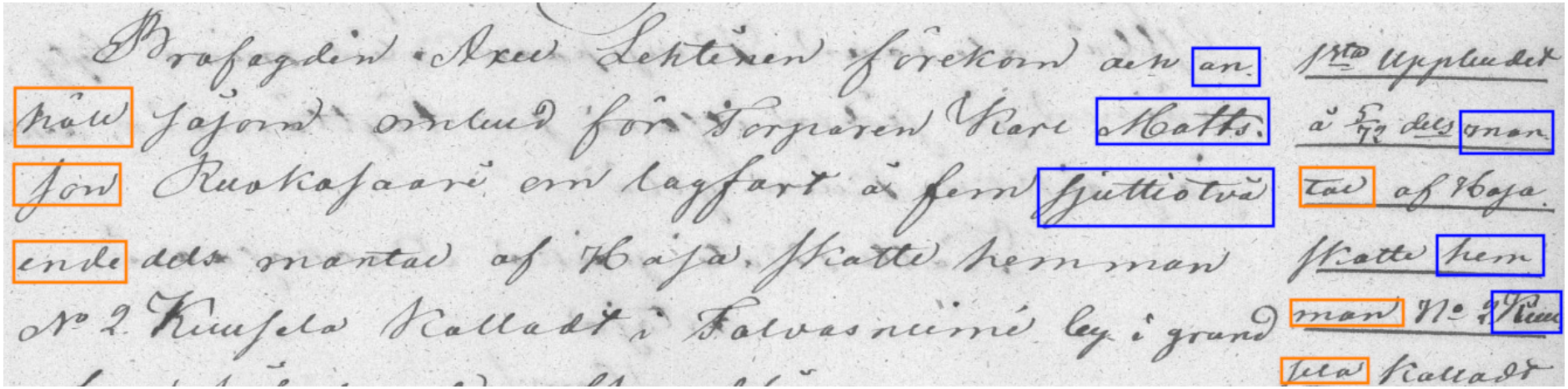

```
# pageID="Bentham-071-021-002-part"
#    keyword confid    bounding box
#
           2  0.929    1  36  20 31
          21  0.064    1  36  24 31
          IT  0.982   33  36  27 31
          IF  0.012   33  36  26 31
     MATTERS  0.989   77  36  99 31
      MATTER  0.011   77  36  93 31
         NOT  0.999  216  36   7 31
     WHETHER  1.000  256  36  99 31
         THE  0.997  389  36  33 31
MIS-SUPPOSAL  1.000  455  36 193 31

         THE  0.927  430  88  30 31
          HE  0.056  434  88  25 31
         ...  ...      ...     ...
```

```
REGARDS  0.857    5 115  84 31
REWARDS  0.138    5 115  90 31
    THE  0.993  110 115  43 31
 MATTER  0.998  160 115  93 31
     OF  0.996  271 115  23 31
   FACT  0.999  306 115  49 31
     OR  0.973  377 115  37 31
     ON  0.021  377 115  42 31
 MATTER  0.990  425 116 100 31
     OF  0.995  542 115  25 31
     BY  0.407  575 115  30 31
    ANY  0.175  575 115  55 31
    ...  ...      ...     ...
    LAW  0.032  575 115  36 31
    LAY  0.031  575 115  55 31
    ...  ...      ...     ...
    PAY  0.012  575 115  59 31
```

```
    THE  0.990    1 198  28 31
 MATTER  0.934   61 198  64 31
     OF  0.988  141 198  28 31
   FAST  0.367  182 198  62 31
    FAR  0.186  182 198  36 31
    ...  ...      ...     ...
   FACT  0.017  182 198  46 31
     AS  0.142  200 198  29 31
    HAS  0.022  200 198  29 31
  WHERE  0.992  255 198  90 31
    YOU  0.761  365 198  45 31
   YOUR  0.030  365 198  47 31
   GOES  0.064  372 198  45 31
SUPPOSE  0.975  429 198 120 31
SUPPOSED 0.024  429 198 125 31
   SOME  0.834  570 198  78 31
 SOONER  0.016  576 198  83 31
    ONE  0.109  580 198  65 31
     ME  0.022  620 198  22 31
```

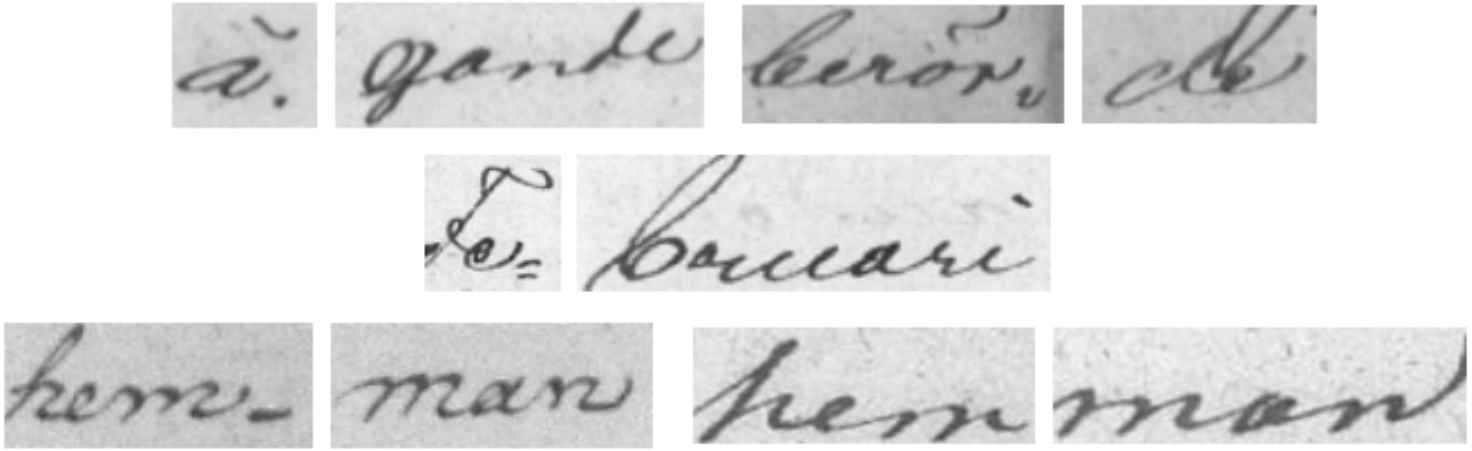
Introduction: Definition



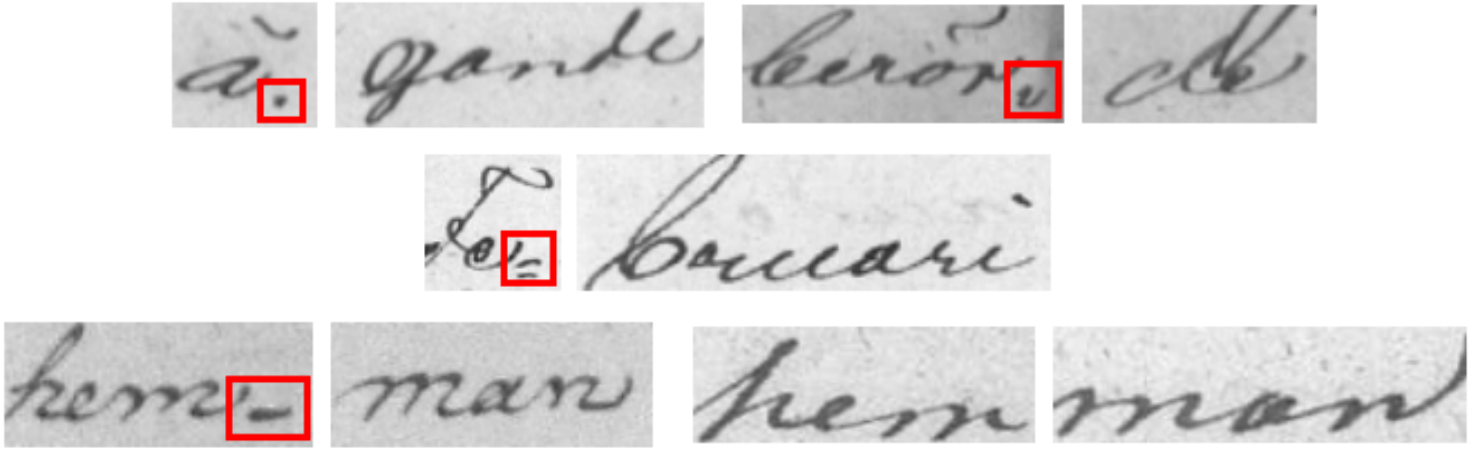
Hyphenated Word: word divided into two HwF's due to a line break.

Typically denoted by adding a special symbol at the end of the prefix.

Introduction: Challenges

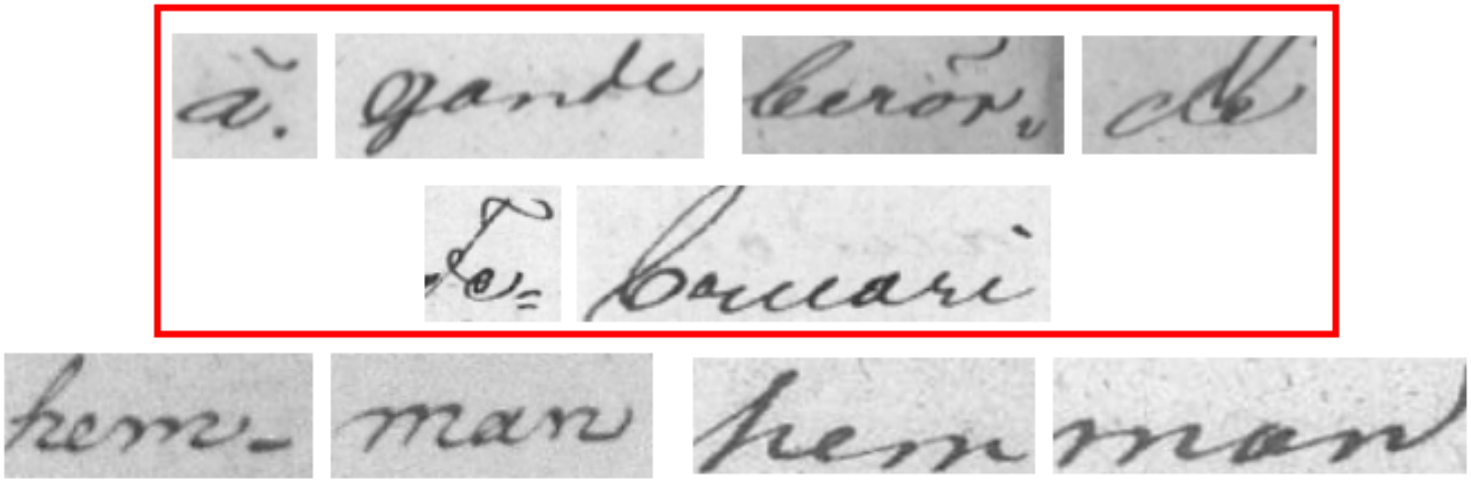


Introduction: Challenges



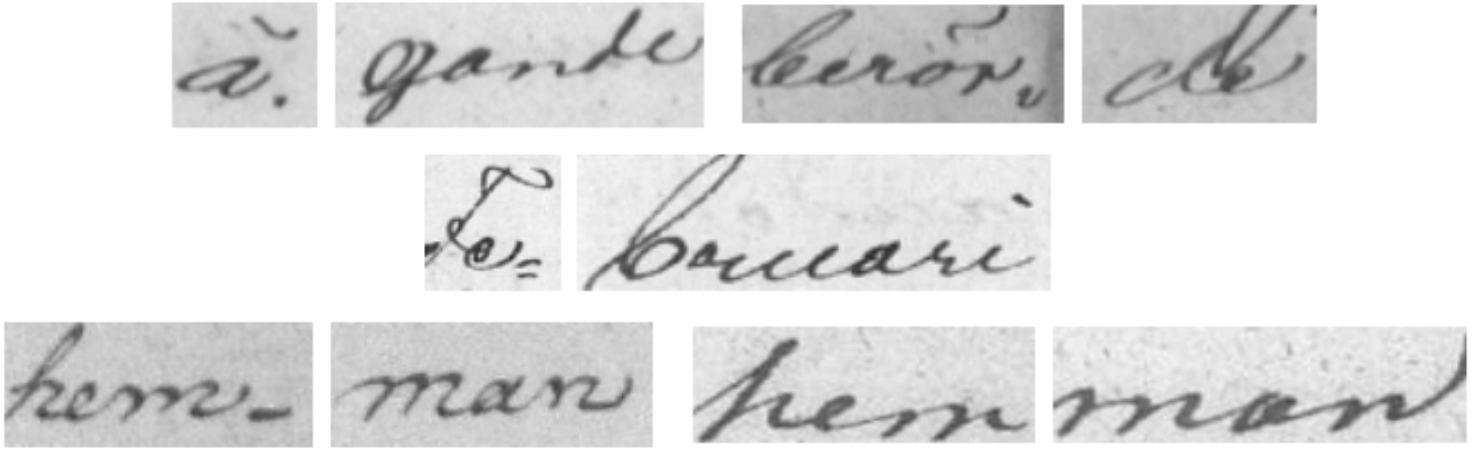
Different hyphenation symbols.

Introduction: Challenges



Might not follow modern hyphenation rules.

Introduction: Challenges



Our aim is IR. We need to retrieve the HypWrd's, not the fragments

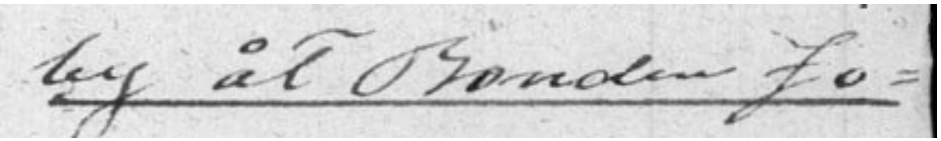
Our approach

Two **offline** phases:

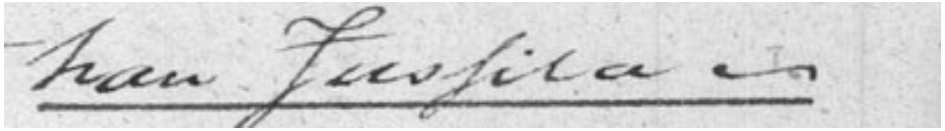
1. Optically predict hyphenated word fragments (HwF's).
2. Join them to form hyphenated words.

Optical and language model

Optical modeling: Prefix and suffix fragments are tagged with “>”



BY AT BONDEN JO>



>HAN JUSSILA

Language modeling: add constraints to n-gram model (hyphenation symbol only at beginning or end of sentence).

Offline merging of HwF's

$$P(R \mid x, b_r, b_s, r, s) \approx \min\left(P(r \mid x, b_r), P(s \mid x, b_s), P(R_h \mid x, b_r, b_s)\right)$$

Where probabilities in red are provided by the PrIx and the ones in blue can be estimated by different methods.

Estimating $P(R_h \mid x, b_r, b_s)$

Plain: Always set to 0.

All combinations: Always set to 1.

Heuristic: 1 *iff* hand-crafted geometric restrictions are fulfilled.

MLP: Estimate it through MLP.

Oracle: 1 *iff*, according to the GT, there are two consecutive textlines beginning and ending with prefix and suffix HwF's, respectively, to which b_r and b_s belong.

Hyphenation to generate HwF queries online

As a baseline, we consider the scenario of using hyphenation software to query hyphenated and not hyphenated words online.

Consider the query "Katarina":

$$\text{Katarina} \vee (\text{Ka}> \wedge >\text{tarina}) \vee (\text{Kata}> \wedge >\text{rina}) \vee (\text{Katari}> \wedge >\text{na})$$

Dataset statistics

Dataset partition:	Dataset		HwF's		
	Train-Val	Test	Train-Val	Test	Overall %
Images	400	200	–	–	–
Lines	25 989	13 341	10 973	5 609	42%
Running words	147 118	73 849	13 081	6 589	9%
Lexicon size	20 710	13 955	4 091	2 677	20%
AllWords query set	–	10 416	–	–	–
MaybeHyph query set	–	1 972	–	–	–

Table: Basic statistics of the FCR-HYP dataset and their hyphenated word fragments (HwF's). All the text has been transliterated and the punctuation marks ignored.

Metrics

To assess IR: mAP and AP.

To assess storage usage: PrIx density.

Results: MaybeHyp queryset

Input Metric	Prlx Pruned by 10^{-5} mAP	AP	density	1-best HTR mAP	AP	density
Plain	0.43	0.80	10	0.35	0.72	1
Pyphen	0.65	0.87	10	0.43	0.74	1
All combin.	0.68	0.88	271	0.44	0.75	2
Heuristic	**0.71**	**0.89**	21	0.45	0.76	1
MLP (10^{-4})	**0.71**	**0.89**	33	**0.46**	**0.77**	1
MLP (0.04)	0.70	0.88	24	0.45	**0.77**	1
MLP (0.35)	0.69	0.88	19	0.45	0.76	1
Oracle	0.71	0.89	12	0.46	0.77	1

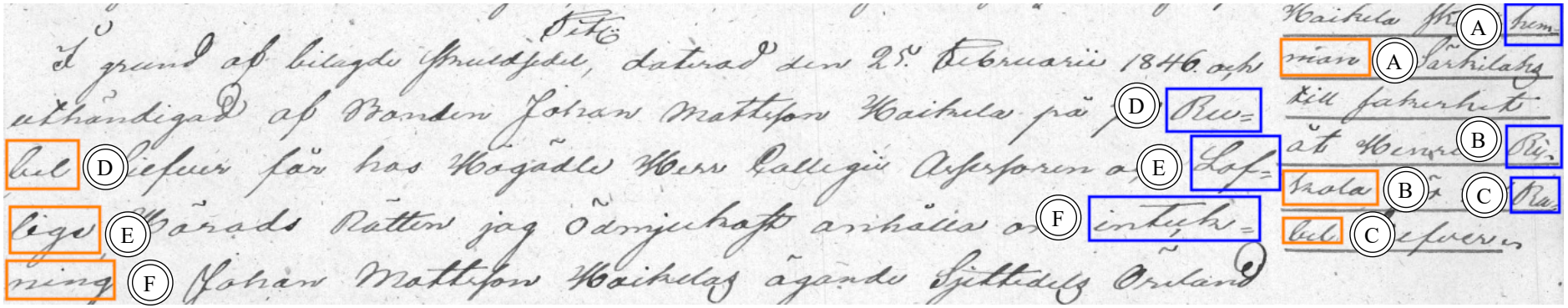
Table: mAP, AP and density with the MaybeHyp queryset.

Results: AllWords queryset

Input Metric	Prlx Pruned by 10^{-5} mAP	AP	density	1-best HTR mAP	AP	density
Plain	0.72	0.83	10	0.46	0.69	1
Pyphen	0.75	0.85	10	0.47	0.69	1
All combin.	0.75	0.85	271	0.47	0.69	2
Heuristic	**0.77**	**0.86**	21	**0.48**	**0.71**	1
MLP (10^{-4})	**0.77**	**0.86**	33	**0.48**	**0.71**	1
MLP (0.04)	0.76	**0.86**	24	**0.48**	**0.71**	1
MLP (0.35)	0.76	**0.86**	19	**0.48**	**0.71**	1
Oracle	0.77	0.86	12	0.48	0.71	1

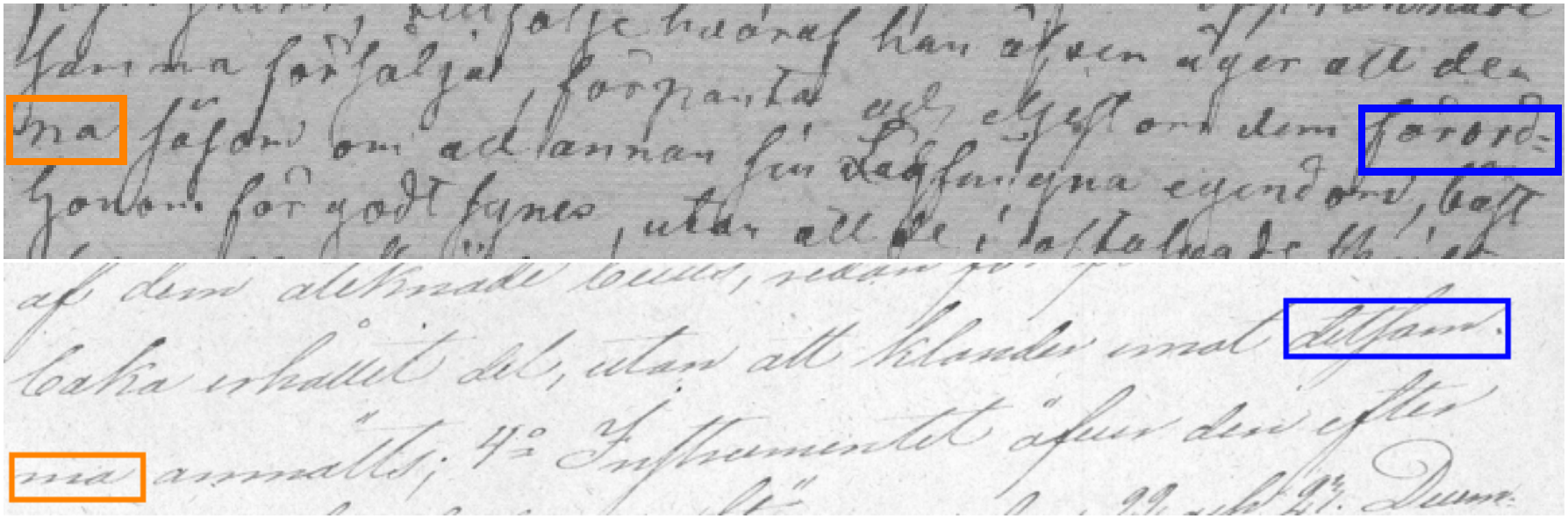
Table: mAP, AP and density with the AllWords queryset.

Illustrative retrieval examples



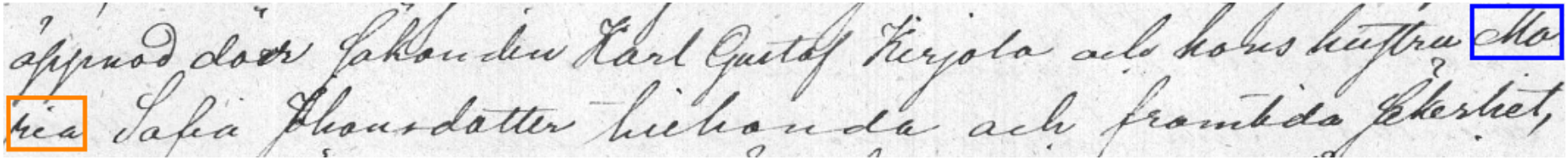
Demo available at:
http://prhlt-carabela.prhlt.upv.es/fcr-hyp-icdar23/

Illustrative retrieval examples



FN's made by the heuristic approach.

Illustrative retrieval examples



FN made by the MLP approach and Pyphen.

Conclusions

- Methods relying on PrIx to allow hyphenated word searches have been developed.
- Heuristic and MLP are the best performing methods.
- There is still room of improvement for the density.

Future works

- Incorporate lexical information in the joining of HwF spots phase.
- Assess these methods using automatic lines.

Search for Hyphenated Words in Probabilistic Indices: a Machine Learning Approach

José Andrés, Alejandro H. Toselli, Enrique Vidal

tranSkriptorium AI, Valencia, Spain

August 21st, 2023

tranSkriptorium

Probabilistic framework

$$P(R \mid x, v) \approx \max_{b \sqsubseteq x} P(R \mid x, v, b) \approx \max_{\substack{b_r, b_s, r, s: \\ rs = v, b_r, b_s \sqsubseteq x}} P(R \mid x, b_r, b_s, r, s)$$

Adapting RP formula to deal with hyphenated instances.

Probabilistic framework

We consider R the conjunction of three boolean random variables: R_r, R_s, R_h.

$$
\begin{aligned}
&P(R_r, R_s, R_h \mid x, b_r, b_s, r, s) \\
&\quad \approx \min\big(P(R_r \mid x, b_r, b_s, r, s), P(R_s \mid x, b_r, b_s, r, s), P(R_h \mid x, b_r, b_s, r, s)\big) \\
&\quad \approx \min\big(P(R_r \mid x, b_r, r), P(R_s \mid x, b_s, s), P(R_h \mid x, b_r, b_s)\big) \\
&\quad \approx \min\big(P(r \mid x, b_r), P(s \mid x, b_s), P(R_h \mid x, b_r, b_s)\big)
\end{aligned}
$$

HTR performance

	no LM	char 8-gram
All tokens	31.1	23.0
Only HwF's	44.0	39.3

Table: WER (in %) with and without LM.

Results

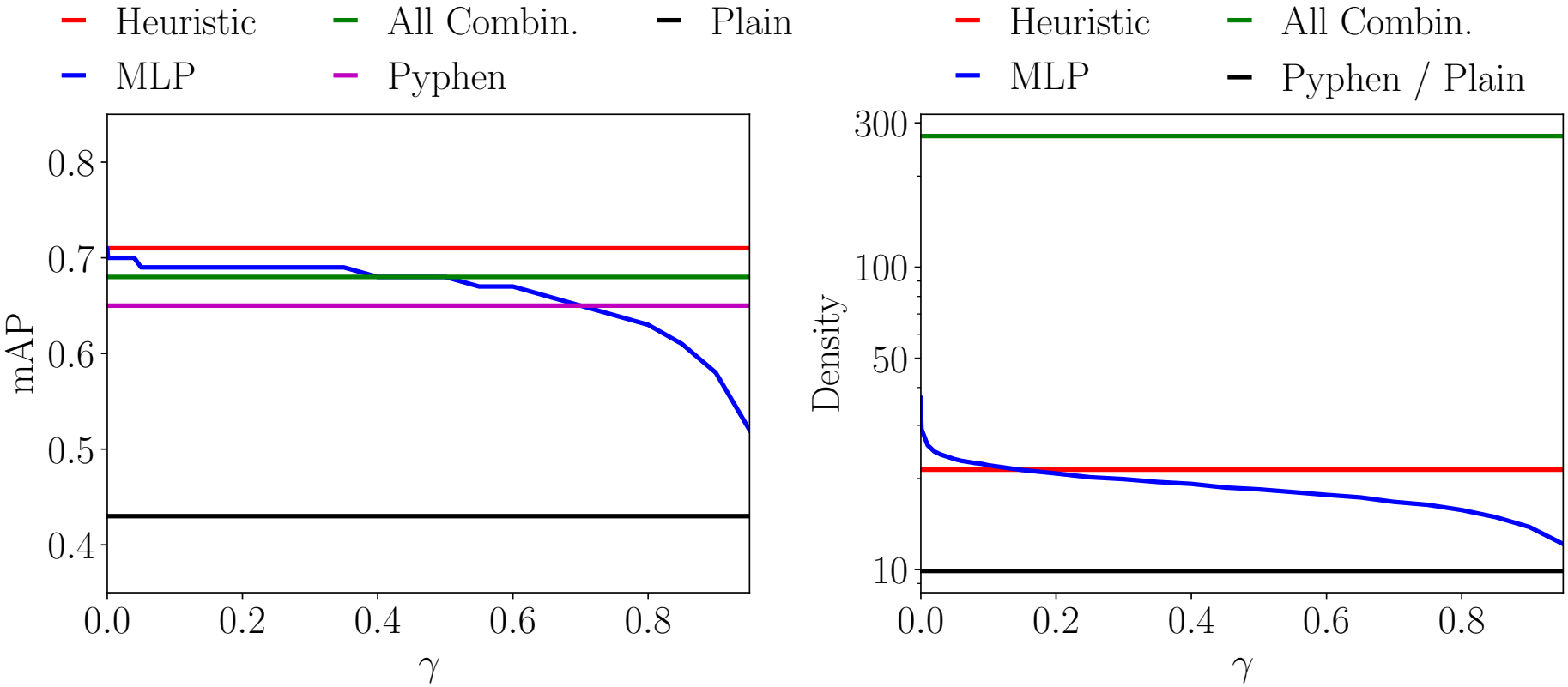


Figure: MaybeHyp mAP (left) and density (right), as a function of the MLP threshold γ, for the different techniques using Prlx's pruned by 10^{-5}.