ICFHR 2018

6th International Conference on Frontiers in Handwriting Recognition

Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records

Eva Lang[†], Joan Puigcerver[‡], Alejandro H. Toselli[‡] and Enrique Vidal[‡]

[†]Archiv des Bistums Passau
Bischoefliches Oridinariat Passau, Passau, Germany

`eva.lang@bistum-passau.de`

[‡]Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain

`{jpuigcerver,ahector,evidal}@prhlt.upv.es`

August 6th, 2018

Outline

Introduction

- Huge amounts of legacy handwritten documents exist, but perhaps more than 99.99% of them are *untranscribed*.

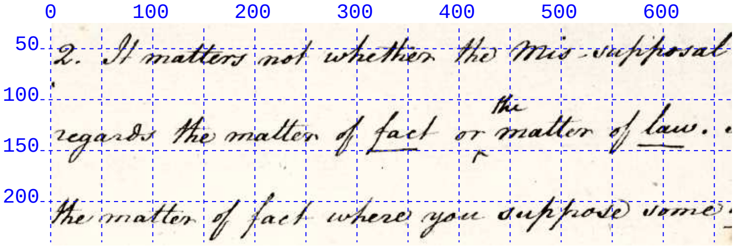
- In particular, text access is in high demand for many *archive* documents: birth, marriage and death records, military draft records, census, property, etc.

  Here we deal with a *German handwritten parish record collection* (16th - 19th c.), held by the Passau Diocesan Archives.

- Rely on Lexicon-free Probabilistic Indices (PI) which allow *fast search & retrieval* and other forms of *text data analysis* from untranscribed handwritten text images.

- Two main contributions of the present work:
  1. Analyze the *impact* of *transliteration* and *PI density* (size) on indexing and search performance.
  2. Successfully explore the *use of PIs* to support structured, multiple-word queries for *information extraction from untranscribed handwritten tables*.

Lexicon-free Probabilistic Index: Example



pageID="Bentham-071-021-002-part"
keyword relPrb bounding box
#

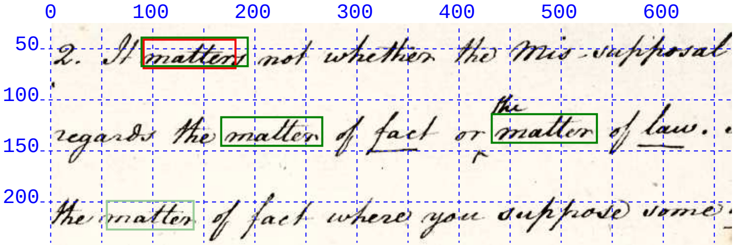
keyword	relPrb			bounding box	
2	0.929	1	36	20	31
21	0.064	1	36	24	31
IT	0.982	33	36	27	31
IF	0.012	33	36	26	31
MATTERS	0.989	77	36	99	31
MATTER	0.011	77	36	93	31
NOT	0.999	216	36	7	31
WHETHER	1.000	256	36	99	31
THE	0.997	389	36	33	31
MIS-SUPPOSAL	1.000	455	36	193	31
THE	0.927	430	88	30	31
LHE	0.056	434	88	25	31
...	...	...		...	

REGARDS	0.857	5	115	84	31
UGARDS	0.138	5	115	80	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
LAM	0.407	575	115	30	31
BIMR	0.175	575	115	55	31
...	...	...		...	
LAW	0.032	575	115	36	31
TAUE	0.031	575	115	55	31
...	...	...		...	
LANE	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
HAE	0.022	200	198	29	31
WHERE	0.992	255	198	90	31
YOU	0.761	365	198	45	31
YOW	0.030	365	198	45	31
GOUS	0.064	372	198	47	31
SUPPOSE	0.975	429	198	120	31
SUPFROSE	0.024	429	198	125	31
SOME	0.834	570	198	78	31
SONER	0.016	576	198	83	31
OME	0.109	580	198	65	31
ME	0.022	620	198	22	31

All character strings or “pseudo-words” which are likely enough to be real words are indexed.

Lexicon-free Probabilistic Index: Example



pageID="Bentham-071-021-002-part"
keyword relPrb 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	**0.998**	**160**	**115**	**93**	**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
LHE	0.056	434	88	25	31
...	...	...		...	

REGARDS	0.857	5	115	84	31
UGARDS	0.138	5	115	80	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
LAM	0.407	575	115	30	31
BIMR	0.175	575	115	55	31
...	...	...		...	
LAW	0.032	575	115	36	31
TAUE	0.031	575	115	55	31
...	...	...		...	
LANE	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
HAE	0.022	200	198	29	31
WHERE	0.992	255	198	90	31
YOU	0.761	365	198	45	31
YOW	0.030	365	198	45	31
GOUS	0.064	372	198	47	31
SUPPOSE	0.975	429	198	120	31
SUPFROSE	0.024	429	198	125	31
SOME	0.834	570	198	78	31
SONER	0.016	576	198	83	31
OME	0.109	580	198	65	31
ME	0.022	620	198	22	31

Spots for **MATTER** and **MATTERS** marked in colors according to their Relevance Probabilities.

From the Filler Model to Lexicon-free Probabilistic Indexing

- Segmentation- & Lexicon-free Filler KWS approaches based on HMM/RNN
  A. Fischer et al., “Lexicon-free handwritten word spotting using character HMMs” Pattern Recognition Letters, 2012.
  V. Frinken et al., “A novel word spotting method based on recurrent neural networks” IEEE TPAMI, 2012.
- Reduce Filler high computing cost using *character lattices* (CL) (same accuracy)
  A. H. Toselli et al., “Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents” ICDAR’13.
- Filler accuracy improved by adding 2-gram character LM (still much slower)
  A. Fischer at al., ”Improving HMM-Based Keyword Spotting with Character Language Models”, ICDAR’13.
- Use 6*-gram LM* to improve Filler accuracy, boost efficiency by means of CLs
  A. H. Toselli et al., “Context-aware lattice based filler approach for key word spotting in handwritten documents”, ICDAR’15.
- *Filler probabilistic interpretation*: leads to correct spotting *Relevance probability*
  Puigcerver et al., “Probab. interpret. and improvements to the HMM-filler for handwritten keyword spotting”, ICDAR’15.
- Further improve accuracy and efficiency of probabilistically interpreted Filler model
  A. H. Toselli et al., “Two methods to improve confidence scores for lexicon-free word spotting in handwritten text” ICFHR’16.
- *Large-scale Lexicon-free Probabilistic Indexing* (PI) based on the probabilistic Filler
  T. Bluche et. al., “Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project” ICDAR’17.

From the Filler Model to Lexicon-free Probabilistic Indexing: Ours

- Segmentation- & Lexicon-free Filler KWS approaches based on HMM/RNN

  A. Fischer et al., “Lexicon-free handwritten word spotting using character HMMs” Pattern Recognition Letters, 2012.

  V. Frinken et al., “A novel word spotting method based on recurrent neural networks” IEEE TPAMI, 2012.

- Reduce Filler high computing cost using *character lattices* (CL) (same accuracy)

  A. H. Toselli et al., “Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents” ICDAR’13.

- Filler accuracy improved by adding 2-gram character LM (still much slower)

  A. Fischer at al., ”Improving HMM-Based Keyword Spotting with Character Language Models”, ICDAR’13.

- Use 6*-gram LM* to improve Filler accuracy, boost efficiency by means of CLs

  A. H. Toselli et al., “Context-aware lattice based filler approach for key word spotting in handwritten documents”, ICDAR’15.

- *Filler probabilistic interpretation*: leads to correct spotting *Relevance probability*

  Puigcerver et al., “Probab. interpret. and improvements to the HMM-filler for handwritten keyword spotting”, ICDAR’15.

- Further improve accuracy and efficiency of probabilistically interpreted Filler model

  A. H. Toselli et al., “Two methods to improve confidence scores for lexicon-free word spotting in handwritten text” ICFHR’16.

- *Large-scale Lexicon-free Probabilistic Indexing* (PI) based on the probabilistic Filler

  T. Bluche et. al., “Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project” ICDAR’17.

Lexicon-free Probabilistic Indexing Search Performance: Impact of Transliteration and Language Modeling

Transliteration: normalize spelling, fold diacritics and case of query strings, etc.

Early: when training char. Optical Models; *Late:* when the Probabilistic Index is built

Average Precision (AP), *mean AP* (mAP) for different transliterations and *language models* (LM)

Transliteration	Character LM	AP	mAP
Early	none	0.70	0.66
Early	3-gram	0.71	0.68
Early	6-gram	0.75	0.69
Late	6-gram	0.69	0.66

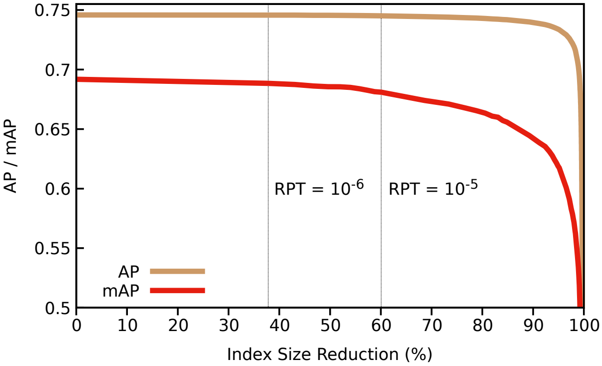
LMs provide useful AP and mAP improvements

Early transliteration proves significantly better

Probabilistic Index Trimming: Effect on Search Performance

Indexing a large number of possibly useless pseudo-word spots do not harm search performance, but do result in large storage overheads, problematic for big collections.

Most unlikely spots can safely be trimmed:



With a Relevance Probability Threshold of 10^{-5}, the average index size per page drops from $56\,953$ to $22\,742$ spots (60%), but mAP falls only from 0.69 to 0.68 (and AP decay is negligeable)

Outline

Information Extraction (IE) from Handwritten Table Images

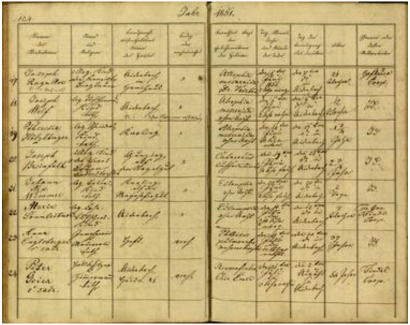
- Handwritten tables perhaps account for more than half of the vast amounts of documents preserved in archives.
- Tables contain important, and often ready-to-use information for many historical studies, such as ethnography, demography, economics, genealogy, etc.
- Accurately transcribing images of handwriting tables is *very difficult:*
  - Ad-hoc, variable, inconsistent and even erratic layouts,
  - difficult line detection,
  - hopeless *reading order* ambiguities,
  - short lines lack linguistic context to help accurate word recognition,
  - etc.

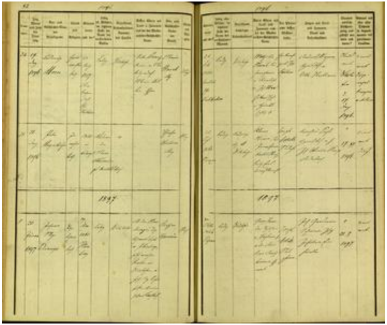
Good news: Probabilistic Indices can support structured, multiple-word queries aiming at complex information extraction from untranscribed images of tabular data.

Towards Information Extraction from Table Images

- From previous works: *PI's support page-level Boolean multi-word queries*
- *PI's hold geometric information* (position, shape and size) of the bounding boxes (BB) of the indexed words
- *Boolean queries*, along with BB-based *geometric reasoning*, can be used to support structured queries for *information extraction from table images*

Example of handwritten table images from the *Passau* dataset:





Structured Multi-word Queries for IE from Table Images

Aim to deal with queries of the form:

⟨**column-heading, column-content**⟩

where **column-heading** is an AND combination of table heading words and **column-content** is a (single) keyword.†

Examples;

- ⟨**ORT, STEINERLEINBACH**⟩ (⟨PLACE, STEINERLEINBACH⟩)
- ⟨**TAUF TAG, APRIL**⟩ (⟨BAPTISM DAY, APRIL⟩)
- ⟨**KRANKHEIT ARZT, FRAISEN**⟩ (⟨CAUSE OF DEATH, SPASMS⟩)
- ⟨**NAMEN DES BRAEUTIGAMS, JOSEF**⟩ (⟨NAME OF THE GROOM, JOSEF⟩)
- ⟨**NAMEN DER BRAUT, MARIA**⟩ (⟨NAME OF THE BRIDE, MARIA⟩)
- ⟨**TAG MONAT JAHR TODES, 1879**⟩ (⟨DAY MONTH YEAR OF DEATH, 1879⟩)

† More complex structured queries can be similarly supported

Probabilistic Framework for Structured Multi-word Query Search

- Let $\mathbf{h} \stackrel{\text{def}}{=} \{h_1, h_2 \ldots, h_I\}$ be the set of column-heading query words and let $R(h_i)$ denote the *relevance probability* (RP) of h_i ($h_i \in \{0,1\}$ relevant Boolean variable).
- Let $s_{i1}, \ldots, s_{iJ_i}$ denote the $J_i \geq 1$ different spots of h_i and $R(s_{ij}) \stackrel{\text{def}}{=} P(R \mid \mathbf{c}_{h_i}, \mathbf{x}_{ij})$ its RP in the image location $\mathbf{x}_{ij}$, where $\mathbf{c}_{h_i}$ is the character spelling of the word h_i.
- Then, the RP of the AND combination for the words in $\mathbf{h}$ is computed as:

$$R(\mathbf{h}) = R(h_1 \wedge h_2 \cdots \wedge h_I) \approx \min_{1 \leq i \leq I} R(h_i) \approx \min_{1 \leq i \leq I} \max_{1 \leq j \leq J_i} R(s_{ij}) \quad \text{(see †)}$$

- Let $v_1, \ldots, v_K$, $K \geq 1$, be the different spots of v retrieved in column locations $\mathbf{x}_1, \ldots, \mathbf{x}_K$ and let $R(v_k) \stackrel{\text{def}}{=} P(R \mid \mathbf{c}_v, \mathbf{x}_k)$ be the RP of the k-th spot.
- The RP of the column-content word v in the considered column is computed as:

$$R(v) \approx \max_{1 \leq k \leq K} R(v_k)$$

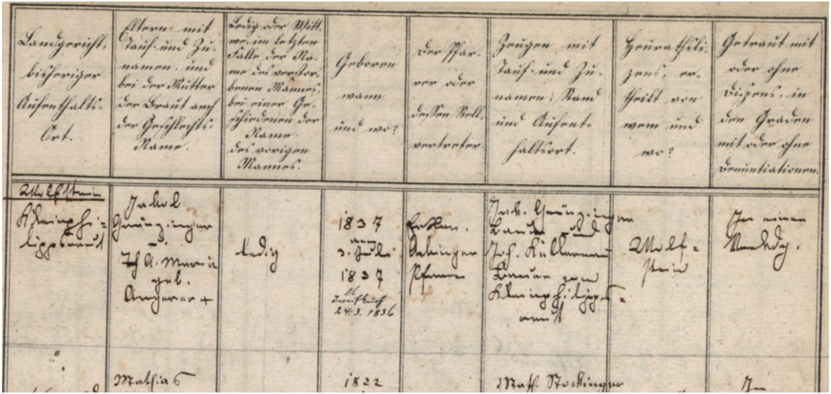
- Finally, the RP of a column-wise structured multi-word query is computed as:

$$R(\langle \mathbf{h}, v \rangle) = R(\mathbf{h} \wedge v) \approx \min\big(R(\mathbf{h}), R(v)\big) \quad \text{(see †)}$$

† A.H. Toselli, E. Vidal, J. Puigcerver and E. Noya-García: Probabilistic Multi-Word Spotting in Handwritten Text Images. To be published in Journal of Pattern Analysis and Applications. 2018.

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩ User query



Example

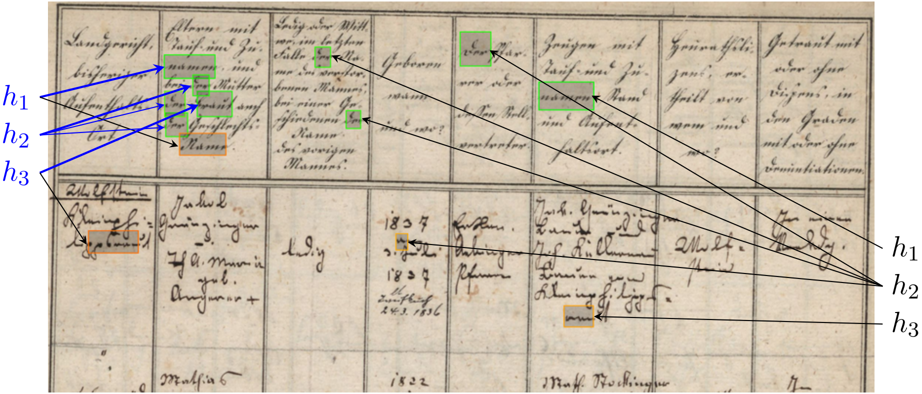
Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

Spotting heading words

h_1 = NAMEN
h_2 = DER
h_3 = BRAUT

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩ Spotting heading words

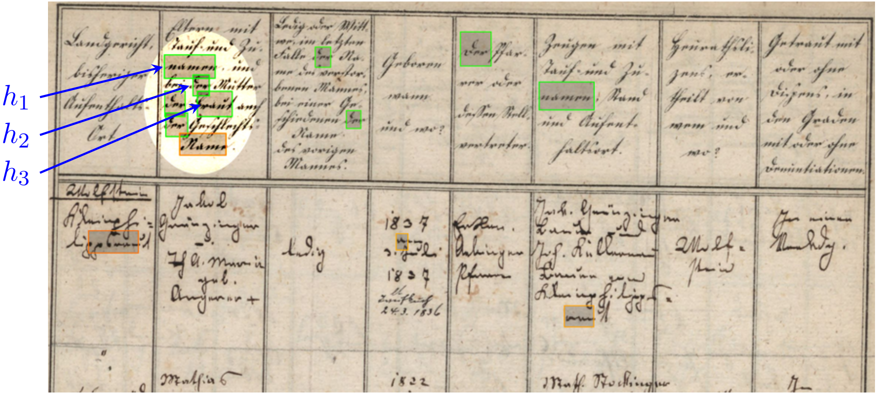


h_1 = NAMEN
h_2 = DER
h_3 = BRAUT

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

Applying geometric restrictions



$h_1 =$ NAMEN
$h_2 =$ DER
$h_3 =$ BRAUT

Relevance Prob.: $R(\mathbf{h}) \approx \min\left(R(h_1), R(h_2), R(h_3)\right)$

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩   Candidate regions for column-content words

Example

Query: ⟨ NAMEN DER BRAUT, MARIA ⟩

Spotting column-content words

$v =$ MARIA

Relevance Prob.: $R(v) \approx \max\left(R(v_1), R(v_2)\right)$

Example

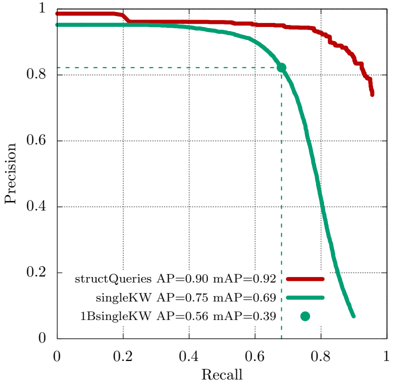
Query: ⟨ NAMEN DER BRAUT, MARIA ⟩ Retrieved spot and its relevance probability

$h_1 =$ NAMEN
$h_2 =$ DER
$h_3 =$ BRAUT
$v =$ MARIA

Relevance Prob.: $R(\langle \mathbf{h}, v \rangle) = R(\mathbf{h} \wedge v) \approx \min\big(R(\mathbf{h}), R(v)\big)$

Information Extraction from Handwritten Table Images: Results

Search performance for single and structured word queries:



Dataset training and test details

- **PASSAU:** German/Latin, *many hands*.
  Training: 200 pages, 102 char CRNN OMs + char 6-gram LM trained on training transcripts; Lexicon: 12 381 tokens.
  Test: 91 page images; *Query set*: 6 500 keywords
- **PASSAUSTRUC:** Table queries in PASSAU.
  Training: same as PASSAU. *Test:* 44 table images; *Query set*: 363 real multi-word structured queries.
- See:
  http://transcriptorium.eu/demots/kws-Passau

Outstanding table information extraction results based on multi-word structured queries

Conclusions

- The present work confirms with another difficult collection the high effectiveness of lexicon-free single-keyword KWS supported by Probabilistic Indices
- PIs have been shown to support structured queries involving many words, which allow for complex information retrieval in text images containing tabular data
- Empirical results validate the proposed approaches for actually indexing the full *Passau* collection, with more than $800\,000$ historical handwritten register images
- A real demonstrator of the indexing and search techniques developed and evaluated in this work is publicly available at:

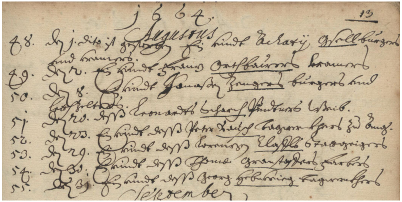
http://transcriptorium.eu/demots/kws-Passau

(no yet supporting table information extraction queries)

Thanks for Your Attention !

PASSAU Collection and Experimental Dataset

XVI-XVIII century collection of historical records. $26,000$ images, written in German.



Information about the baptized, married and die parishioners of the various Passau's Diocese parishes.

A small dataset of 291 images of varied types produced by ABP with GT transcripts and detected text lines.

	Train+Val	Test	TabTest
#Pages	200	89	44
#Lines	29 314	16 376	11 710
RWs	72 848	37 354	21 027
RWs w/o PMs	–	26 709	15 141
Lex-size	12 381	6 532	3 455
#Chars	220	187	119
Transl. Lex-size*	11 160	5 801	3 141
#Transl. Chars*	99	87	73

Statistics for single and structured word queries.

Complex, varied layout, many tables, etc.: 2.4 average words/line ⇒ low LM impact.

* *Transliteration:* all chars uppercase, no diacritics, non-ASCII symbols mapped to ASCII "equivalents"

Probabilistic Index Size and Transliteration

Probabilistic Indices:

- may become huge (large amounts of storage) for vast manuscript collections.
- contain large quantities of pseudo-words which probably will never be spotted.

Solution: PI size reduction through filtering out entries whose relevance probability scores fall bellow an specified threshold.

The medieval German record collection used here contains:

- different spelling variations of the same word (e.g., accents, umlauts, tie bar, . . .)
- 263 UTF-8 different symbols, most of which are/contain non-ASCII characters.
- most of such characters can not be typed on standard keyboards.

Solution: Every char/symbol is transliterated by case folding and by removing diacritics and mapping non-ASCII symbols onto their ASCII equivalents.

Remov. Diacrit.		Non-ASCII to ASCII					
ċ, č, č́	C	Æ, æ	AE	ĳ	II	ŋ	EN
è, ē, ë	E	Œ, œ	OE	ß	SS	ꝰ	US
m̄, m̌, ṃ	M	p̖, ṕ	PRO	đ	DE	ẟ	DER

The benefits are two-fold: a) simplify the composition of queries and b) avoid the waste of probability mass which often leads to degrade search performance.

PI performance: Impact of Transliteration and Language Modeling

Transliteration: normalize spelling and fold diacritics and case of query strings.
Early: at the of optical modelling; *Late:* after the PI is built

Transliteration	Latt-type	Char LM	AP	mAP	$MxRc_{10}$
Early	CLs	none	0.701	0.661	0.861
Early	CLs	3-gram	0.712	0.677	0.876
Early	**CLs**	**6-gram**	**0.746**	**0.692**	**0.886**
Late	CLs	6-gram	0.692	0.662	0.854
Early	1-best	6-gram	0.559	0.387	0.680
Late	1-best	6-gram	0.492	0.331	0.613

Average Precision (AP), *mean AP* (mAP) and *maximum recall at 10% precision* ($MxRc_{10}$) for different *character lattices* and *language models* (LM).

Average Precision (AP) versus Mean Average Precision (mAP)

	AP	**mAP**
Rank type	Global	Local
Averaging type	Micro: over all query events	Macro: over the APs of isolated queries
Score consistency impact	yes	no
Demanding relevant queries	not for all	yes for all
Invariant to monotonic transformation	no	yes