How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures

Mohamed Khemakhem, Ioana Galleron, Geoffrey Williams, Laurent Romary, Pedro Javier Ortiz Suárez

How OCR Performance can Impact on the Automatic Extraction of Dictionary Content Structures

Mohamed Khemakhem[1,2,3], Ioana Galleron[4], Geoffrey Williams[6], Laurent Romary[1], Pedro Ortiz Suárez[1,5]

1. **Inria-ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities**
2. **UPD7 - Université Paris Diderot - Paris 7**
3. **CMB - Centre Marc Bloch, Berlin**
4. **Université Sorbonne-Nouvelle**
5. **Sorbonne Université**
6. **Université Grenoble Alpes**

In the last decade, OCR progress has triggered a massive trend towards the digitisation of legacy documents, with several Digital Humanities projects[1][2][3] exploring means for structuring retro-digitised dictionaries. However there is a lack of awareness of the impact of the OCRs quality on the information extraction process. In this work, we shed light on the relationship between these two steps through experiments carried out with a TEI-based system for automatic parsing of dictionaries.

Our work concerns “the Basnage”, a complex dictionary resulting from the complete revision and enlargement in 1701 of the ‘Dictionnaire Universel’ of Abbé Furetière, initially published in 1690. In order to obtain an XML/TEI version of this work, we use GROBID-Dictionaries [1,2], a machine learning system for cascade parsing and extraction of TEI structure in dictionaries. The tool’s models have been tested on different categories of entry based documents with lexical and encyclopedic content.
We used two differently OCRied versions of the first volume of the Basnage[4] following the process described in an earlier experiment [3] which relies on the power of iterative training of HTR models of Transkribus[5] framework:

- **Sample 1:** created using an HTR model trained with 28 pages and a low image quality document

[1] https://basnage.hypotheses.org/
[2] https://elex.is/
[3] http://nenufar.huma-num.fr/presentation/
[4] https://archive.org/details/b30455376_0001/page/n27
[5] https://transkribus.eu/Transkribus/

- **Sample 2:** created using an HTR model trained with 108 pages and high image quality document

The results of our experiment are as follows:

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<ab>	99.95	99.95	99.75	81.48	73.33	77.19
<fw type="footer">	100	76.47	86.67	84.62	91.67	88
<fw type="header">	92.59	80.65	86.21	100	90	94.74

Table 1: Field Level Evaluation of the Dictionary Segmentation Model

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<dictScrap>	81.82	85.71	83.72	100	90	94.74
<entry>	85.85	80.53	83.11	89.47	91.07	90.27
<pc>	92.59	96.15	94.34	93.75	97.56	95.62

Table 2: Field Level Evaluation of the Dictionary Body Segmentation Model

	Sample 1			*Sample 2*		
TEI element	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
<etym>	87.5	60	71.19	73.68	71.79	72.73
<form>	94.44	92.73	93.58	92.24	96.4	94.27
<pc>	90.91	69.44	78.74	88.97	80.13	84.32
<re>	33.33	9.09	14.29	55.56	22.73	32.26
<sense>	67.65	59.28	63.19	77	76.65	76.84
<xr>	100	80	88.89	100	100	100

Table 3: Field Level Evaluation of the Lexical Entry Model

Conclusion

Two main conclusions could be drawn from these experiments. First, the OCRisation process has an important impact on the performance of the automatic parsing of TEI structures. The impact becomes more significant for extracting more granular information. Second, the models implemented in GROBID-Dictionaries are resistant to the noise introduced by the OCR system opening perspectives for exploiting more available digitised material.

References

1. Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *electronic lexicography, eLex 2017*, Sep 2017, Leiden, Netherlands.
2. Mohamed Khemakhem, Axel Herold, Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX workshop at LREC 2018*, May 2018, Miyazaki, Japan. 2018.
3. David Lindemann, Mohamed Khemakhem, Laurent Romary. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. *European Association for Digital Humanities (EADH) Conference*, EADH, Dec 2018, Galway, Ireland.

Biographies

Mohamed Khemakhem is a PhD candidate at Inria, team ALMAnaCH (Paris), Paris 7 University and Centre Marc Bloch (Berlin). His research is focused on parsing lexical and encyclopedic legacy resources using standard-based machine learning models.

Ioana Galleron is a professor of French literature and Digital Humanities at Sorbonne-Nouvelle and UMR 8094 LATTICE of CNRS. She works on computer assisted literary analysis.

Geoffrey Williams is a Professor of Applied Linguistics at the University of South Brittany and researcher at UMR 5316, Litt & Arts at the University Grenoble Alpes. He is a e-lexicographer and leads the ANR BasNum project.

Laurent Romary is senior researcher at Inria, team ALMAnaCH and works on data modelling and standards in humanities computing.

Pedro Ortiz Suárez is a PhD candidate at Inria, team ALMAnaCH (Paris) and Sorbonne Université. His research is focused on enriching lexical and encyclopedic legacy resources using deep learning models.