Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields

Mohamed Khemakhem, Luca Foppiano, Laurent Romary

▶ **To cite this version:**

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. hal-01508868v1

HAL Id: hal-01508868

https://hal.science/hal-01508868v1

Submitted on 17 Aug 2017 (v1), last revised 29 Aug 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields

Mohamed Khemakhem[1,2,4], Luca Foppiano[1] and Laurent Romary[1,3]

[1]Inria - ALPAGE, Paris
[2]Centre Marc Bloch, Berlin
[3]Berlin-Brandenburg Academy of Sciences and Humanities, Berlin
[4]Université Paris-Diderot, Paris
{mohamed.khemakhem,luca.foppiano,laurent.romary}@inria.fr

January 12, 2017

An important number of digitized lexical resources remain unexploited due to their unstructured content. Manually structuring such resources is a costly task given their multifold complexity. Our goal is to find an approach to automatically structure digitized dictionaries, independently from the language or the lexicographic school or style. In this paper we present a first version of **GROBID-Dictionaries**[1], an open source machine learning system for lexical information extraction.

Our approach is twofold: we perform a cascading structure extraction, while we select at each level specific features for training.

We followed a "divide to conquer" strategy to dismantle text constructs in a digitized dictionary, based on the observation of their layout. Main pages (see Figure 1) in almost any dictionary share three blocks: a header (green), a footer (blue) and a body (orange). The body is, in its turn, constituted by several entries (red). Each lexical entry can be further decomposed (see Figure 2) as: form (green), etymology (blue), sense (red) or/and related entry. The same logic could be applied further for each extracted block but in the scope of this paper we focus just on the first three levels.

The cascading approach ensures a better understanding of the learning process's output and consequently simplifies the feature selection process. Limited exclusive text blocks per level helps significantly in diagnosing the cause of prediction errors. It allows an early detection and replacement of irrelevant selected features that can bias a trained model. In such a segmentation, it

Figure 1: First and second segmentation levels of a dictionary page [5]

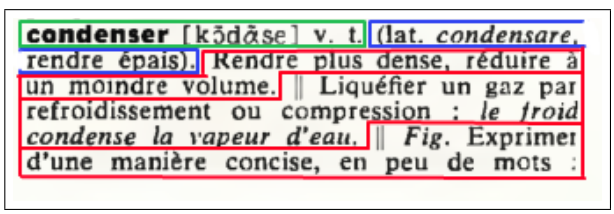

Figure 2: Third segmentation level of a lexical entry [5]

[1]Available at https://github.com/MedKhem/grobid-dictionaries under the Apache License.

becomes more straightforward to notice that, for instance, the token position in the page is very relevant to detect headers and footers and has almost no pertinence for capturing a sense in a lexical entry which is very often split on two pages.

To implement our approach, we took up the available infrastructure from GROBID [7], a machine learning system for the extraction of bibliographic metadata. GROBID adopts the same cascading approach and uses Conditional Random Fields (**CRF**) [6] to label text sequences. The output of Grobid dictionary is planned to generate a TEI compliant encoding [2, 9] where the various segmentation levels are associated with an appropriate XML tessellation. Collaboration with COST ENeL are ongoing to ensure maximal compatibility with existing dictionary projects.

Our experiments justify so far our choices, where models for the first two levels trained on two different dictionary samples have given a high precision and recall with a small amount of annotated data. Relying mainly on the text layout, we tried to diversify the selected features for each model, on the token and line levels. We are working on tuning features and annotating more data to maintain the good results with new samples and to improve the third segmentation level.

While just few task specific attempts [1] have been using machine learning in this research direction, the landscape remains dominated by rule based techniquess [4, 3, 8] which are ad-hoc and costly, even impossible, to adapt for new lexical resources.

References

[1] Petra Bago and Nikola Ljubešić. Using machine learning for language and structure annotation in an 18th century dictionary. In *Electronic lexicography in the 21st century: linking lexical data in the digital age*, 2015.

[2] Gerhard Budin, Stefan Majewski, and Karlheinz Mörth. Creating lexical resources in tei p5. a schema for multi-purpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3), 2012.

[3] Diaa Mohamed Fayed1 Aly Aly Fahmy and Mohsen Abdelrazek Rashwan3 Wafaa Kamel Fayed. Towards structuring an arabic-english machine-readable dictionary using parsing expression grammars. 2014.

[4] Aida Khemakhem, Imen Elleuch, Bilel Gargouri, and Abdelmajid Ben Hamadou. Towards an automatic conversion approach of editorial arabic dictionaries into lmf-iso 24613 standardized model. 2009.

[5] Larousse. *Dictionnaire encyclopédique pour tous (tome I et II) I Dictionnaire des noms communs en couleurs, II Dicionnaire des noms propres en couleurs Collectif*, page 208. France loisirs, 1975.

[6] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513. Association for Computational Linguistics, 2010.

[7] Patrice Lopez and Laurent Romary. Grobid - information extraction from scientific publications. *ERCIM News*, 2015, 2015.

[8] Agnieszka Mykowiecka, Piotr Rychlik, and Jakub Waszczuk. Building an electronic dictionary of old polish on the base of the paper resource. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage at LREC*, pages 16–21, 2012.

[9] Laurent Romary and Werner Wegstein. Consistent modelling of heterogeneous lexical structures. *Journal of the Text Encoding Initiative*, 2012.