Scaling up Automatic Structuring of Manuscript Sales Catalogues

Lucie Rondeau Du Noyer, Simon Gabay, Mohamed Khemakhem, Laurent Romary

Scaling up Automatic Structuring of Manuscript Sales Catalogues

Lucie Rondeau du Noyer
{surname.name@chartes.psl.eu}
Ecole des Chartes, Paris
Simon Gabay
{surname.name@unine.ch}
Université de Neuchâtel
Mohamed Khemakhem
{name.surname@inria.fr}
Inria, team ALMAnaCH, Paris
Centre Marc Bloch, Berlin
Université Paris Diderot, Paris
Laurent Romary
{name.surname@inria.fr}
Inria, team ALMAnaCH

Keywords: Machine learning, manuscript sales catalogues, 19th c. France

Manuscript Sales Catalogues (MSC) are highly important for authenticating documents and studying the reception of authors. Their regular publication throughout Europe since the beginning of the 19th c. has consequently raised the interest around scaling up the means for automatically structuring their contents.
Following successful first encoding tests with *GROBID-Dictionaries* [1,2] on a single MSC collection [3], we aim in this paper to present the results of more advanced tests of the system's capacity to handle a larger corpus with MSC of different dealers, and therefore multiple layouts.

Corpus

Four different types of catalogues published between the middle of the 19th c. and the beginning of the 20th c. have been tested.

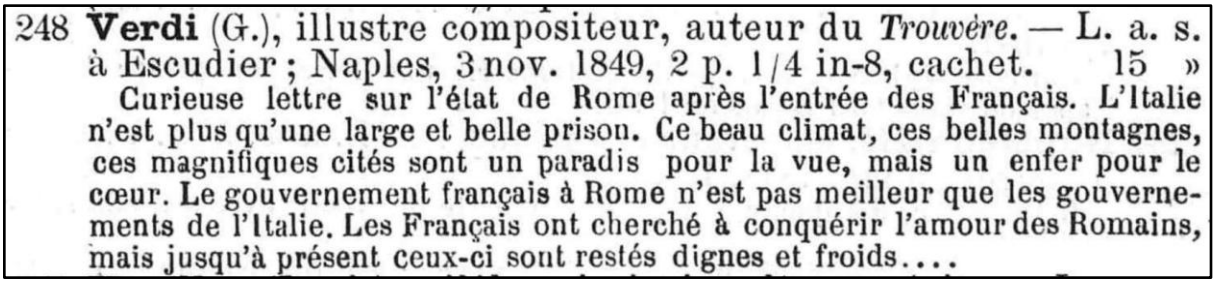
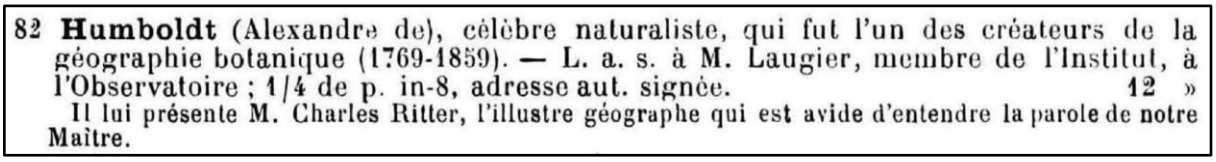
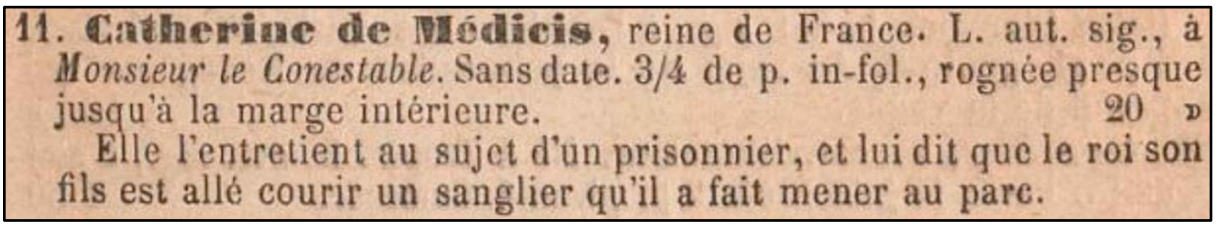

248 **Verdi** (G.), illustre compositeur, auteur du *Trouvère*. — L. a. s. à Escudier ; Naples, 3 nov. 1849, 2 p. 1/4 in-8, cachet. 15 »
Curieuse lettre sur l'état de Rome après l'entrée des Français. L'Italie n'est plus qu'une large et belle prison. Ce beau climat, ces belles montagnes, ces magnifiques cités sont un paradis pour la vue, mais un enfer pour le cœur. Le gouvernement français à Rome n'est pas meilleur que les gouvernements de l'Italie. Les Français ont cherché à conquérir l'amour des Romains, mais jusqu'à présent ceux-ci sont restés dignes et froids....

Figure 1 - Type 1: *Revue des autographes*, Gabriel Charavay. (Première série N°42, Decembre 1874)



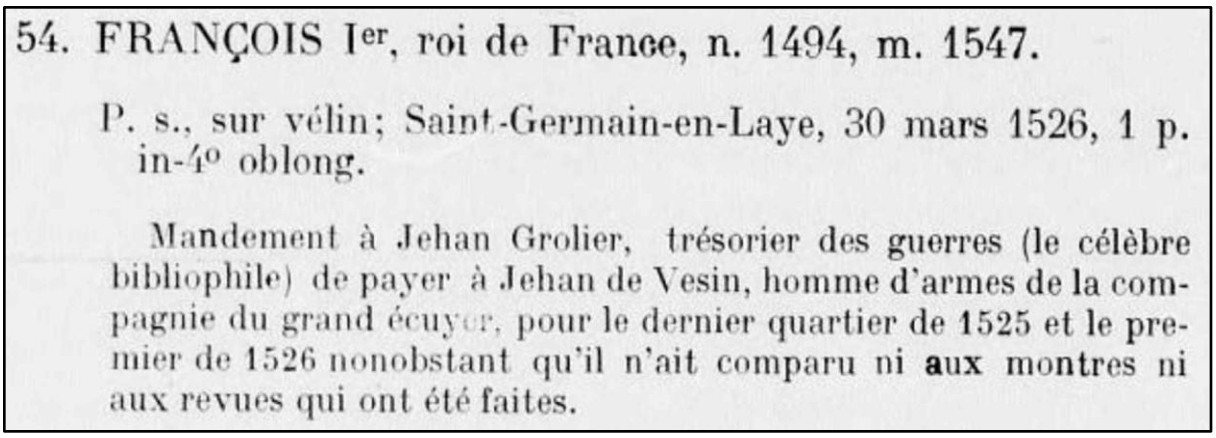
82 **Humboldt** (Alexandre de), célèbre naturaliste, qui fut l'un des créateurs de la géographie botanique (1769-1859). — L. a. s. à M. Laugier, membre de l'Institut, à l'Observatoire ; 1/4 de p. in-8, adresse aut. signée. 12 »
Il lui présente M. Charles Ritter, l'illustre géographe qui est avide d'entendre la parole de notre Maître.

Figure 2 - Type 2: *Revue des autographes, des curiosités de l'histoire et de la biographie*, Gabrielle Charavay (Seconde série N°56, 1934)



11. **Catherine de Médicis,** reine de France. L. aut. sig., à *Monsieur le Conestable*. Sans date. 3/4 de p. in-fol., rognée presque jusqu'à la marge intérieure. 20 »
Elle l'entretient au sujet d'un prisonnier, et lui dit que le roi son fils est allé courir un sanglier qu'il a fait mener au parc.

Figure 3 - Type 3: *Catalogue de lettres autographes et manuscrits*, Auguste Laverdet (N°1, April 1856.)



54. FRANÇOIS Ier, roi de France, n. 1494, m. 1547.

P. s., sur vélin; Saint-Germain-en-Laye, 30 mars 1526, 1 p. in-4° oblong.

Mandement à Jehan Grolier, trésorier des guerres (le célèbre bibliophile) de payer à Jehan de Vesin, homme d'armes de la compagnie du grand écuyer, pour le dernier quartier de 1525 et le premier de 1526 nonobstant qu'il n'ait comparu ni aux montres ni aux revues qui ont été faites.

Figure 4 - Type 4: *Catalogue d'une intéressante collection de lettres autographes*…, Etienne Charavay (December, 14th 1908)

Experiment

To parse the presented MSC corpus we followed the same encoding presented in earlier experiments [3]. We tried then to focus our experiments on two aspects: feature engineering and cumulative samples training.

For the former we tested tuning the GROBID models at three levels of segmentation following two variations: unigram and bigram features. The difference between the two categories is that a label predicted by a trained model is based only on the features of the input token - case of unigram - where a bigram feature template takes also the label of the previous token into consideration.

For the second experimenting aspect we tested the performance of the system on models trained separately for each layout and a general model with all the data.

To that end we used 10 annotated pages for training and 5 others for evaluation, chosen to be representative of each series of catalogues. [4]

	GROBID Models		
MSC	**Lexical Entry**	**Form**	**Sense**
Type 1	72.49	93.72	73.62
Type 2	57.07	84.42	60.5
Type 3	60.07	74.71	48.07
Type 4	60.58	92.03	40.91
All types mixed	63.99	86.42	54.81

Table 1: All fields F1-score of **Unigram Feature Templates**

	GROBID Models		
MSC	**Lexical Entry**	**Form**	**Sense**
Type 1	98.05	100	98.71
Type 2	99.16	95.02	90.78
Type 3	96.01	92.89	88.1
Type 4	92.78	96.83	86.64
All types mixed	95.43	97.44	92.77

Table 2: All fields F1-score of **Bigram Feature Templates**

Conclusion

Two important conclusions can be drawn from this test. First, bigram feature templates are more efficient than unigram templates. Second, a general model potentially increases scores for certain levels (*form* and *sense*) but not all of them (*lexical entry*), which raises the question of the pertinence of a hybrid model, choosing the best solution for each level.

References

1. Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, et *al*.. Automatically Encoding Encyclopedic-like Resources in TEI. *The annual TEI Conference and Members Meeting*, Sep 2018, Tokyo, Japan.

2. 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.

3. Mohamed Khemakhem, Axel Herold, Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan. 2018.

4. Lucie Rondeau du Noyer, Simon Gabay, Mohamed Khemakhem, Laurent Romary. *Training and evaluation data for encoding Manuscript Sales Catalogues with GROBID dictionaries*, Paris: École nationale des chartes (PSL)/Neuchâtel: université de Neuchâtel, 2019, https://github.com/lairaines/grobid_TEI_2019.

Biographies

Lucie Rondeau du Noyer is a " History and New Technologies" masters' student at the Ecole Nationale des Chartes and a graduate student at the Ecole Normale supérieure (Paris).

Simon Gabay is post-doc at the University of Neuchâtel (Switzerland), where he teaches DH and carries research on 17th c. French literature and modern manuscripts. He is currently working on a database of sold manuscripts in 19th c. France.

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.

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