Transfert de connaissance pour l'analyse des registres manuscrits de la Comédie-Italienne au XVIIIe siècle : de l'image au langage

Granet Adeline[1], Roman-Jimenez Geoffrey[1], Rubellin Françoise[2], Quiniou Solen[1], Morin Emmanuel[1], Mouchère Harold[1], Viard-Gaudin Christian[1]
E-mail: prénom.nom@univ-nantes.fr

1. UMR 6004, LS2N, Université de Nantes, 44000, Nantes, France.
2. EA , 4276, L'AMO, Université de Nantes, 44000, Nantes, France.

Les connaissances disponibles sur le développement de la Comédie Italienne (CI) et du théâtre de La Foire au XVIIIe siècle sont très limitées. A travers le projet ANR CIRESFI[1], des chercheurs en Sciences Humaines et Sociales ont comme ambition de mettre à jour des connaissances nouvelles relatives à l'intégration des comédiens italiens en cette période où le contexte politique était résolument hostile.
Dans cette optique, un ensemble de registres comptables de la CI, disponible à la BNF[2], a été numérisé. Cette ressource représente environ 28 000 images, parmi lesquelles des pages de comptes journaliers, mensuels ou annuels, des notes volantes, la liste des pensionnaires, ainsi que les couvertures des registres et des pages blanches. La consultation manuelle de cette masse de données demande beaucoup de temps, même pour des spécialistes du domaine. Pour faciliter la recherche d'information, nous proposons la mise en place d'un système d'extraction automatique des zones d'information.

Le processus d'extraction d'information implique plusieurs étapes consécutives : la segmentation automatique des images en blocs, lignes ou mots suivant la tâche que l'on souhaite réaliser ; suivie de leur transcription par des outils de reconnaissance d'écriture. Les méthodes dites "automatiques" impliquent généralement de pouvoir réaliser un apprentissage et des tests à partir d'images manuellement annotées de la collection. Or, dans notre étude cette vérité n'est pour l'instant pas disponible. Pour résoudre ce problème, nous proposons deux stratégies :

- la première consiste en un site d'annotation collaborative dans lequel les utilisateurs peuvent réaliser trois tâches distinctes : identifier les zones d'informations, transcrire les informations et vérifier les propositions de transcription[3];
- la seconde stratégie, celle que nous présentons ici, est basée sur un apprentissage par transfert de connaissances. Dans ce cas, le système est entraîné sur des images différentes du corpus cible, mais pour lesquelles la transcription recherchée est disponible.

Parmi toutes les informations contenues dans les registres, nous ciblons les titres de pièces répertoriés pour chaque représentation théâtrale. En effet la majorité des pages sont des registres quotidiens qui commencent par lister les pièces jouées ce jour. Nous disposons d'une liste de toutes ces pièces qui ont pu être rassemblées lors des études précédemment menées en SHS [1]. Il convient toutefois de noter que cette liste est fortement bruitée. En particulier, un même titre peut se décliner en de nombreuses variantes, intégrant de possibles abréviations.

[1] Contrainte et Intégration : pour une Ré-Évaluation des Spectacles Forains et Italiens sous l'ancien régime. http://cethefi.org/ciresfi.html
[2] Bibliothèque Nationale de France, exemple d'unregistre sur Gallica : http://gallica.bnf.fr/ark:/12148/btv1b52501115q
[3] Site de Crowdsourcing : http://recital.univ-nantes.fr/

Sur chaque page, la détection de la zone de titre est réalisée par l'utilisation de la méthode DMOS [2] dédiée à la reconnaissance de la structure des documents. La séparation entre les lignes de textes est effectuée à l'aide de l'algorithme *Seam Carving* proposé par Arvanitopoulos *et al* [4].

La transcription des lignes de titres est réalisée à partir d'un réseau de neurones récurrent profond et d'un étiquetage par classification temporelle (BLSTM-CTC [3]). Ce système nécessite une importante quantité de données d'apprentissage. Nous avons sélectionné trois ensembles de données étiquetés :

- Georges Washington (GW) [5] : des lettres de correspondances en anglais ;
- Los Esposalles (ESP) [6] : des registres de mariages espagnols ;
- RIMES [7] : des demandes administratives en français.

Les deux premiers sont des documents datant du XVIIIe siècle mais de natures très différentes ; tandis que la dernière ressource est contemporaine. Les documents de la même période que nos documents de la CI, vont être utiles pour la reconnaissance de certains caractères qui ont évolués avec le temps comme la forme longue du "s", ainsi que le style de l'écriture très caractéristique de cette époque. Il est important également d'avoir une ressource en français, car le BLSTM-CTC apprend implicitement un modèle de langage.

Le système BLSTM-CTC fournit en sortie une matrice de longueur T, qui correspond à la longueur de la trame construite à partir de l'image, et de largeur 75, ce qui correspond aux 74 caractères ciblés et un caractère joker, appelé *blank*. Nous utilisons deux méthodes diamétralement opposées pour décoder cette sortie. La première, appelée *Best Path*, consiste à retenir le caractère le plus actif à chaque instant. Les caractères *blank* sont ensuite supprimés de la séquence obtenue, ainsi que les caractères identiques consécutifs non séparés par un *blank*. Finalement, une distance d'édition est calculée entre la transcription candidate obtenue en sortie et la référence candidate. Les titres avec la distance la plus courte sont conservés. La seconde méthode est beaucoup plus restrictive que la précédente car l'alignement est forcé entre une référence candidate et la matrice de sortie. Pour cela, toutes les références candidates provenant d'une liste sont évaluées sur la matrice, et seules les références fournissant un des dix meilleurs scores comme transcription candidate, sont conservées.
Nous pouvons évaluer notre système facilement sur les bases de données RIMES, GW et ESP en adoptant les répartitions fournies en apprentissage, validation et test. Pour la CI, nous avons constitué une liste de plus de 2 000 titres de pièces, et nous avons également, transcrit manuellement 150 lignes d'images.

Dans cette première partie de projet, notre but est de transcrire le plus fidèlement possible les titres tels qu'ils sont écrits dans les registres. Cependant, plusieurs difficultés résident encore pour assurer l'alignement car les titres dans les images sont parfois abrégés fortement ou coupés sur plusieurs lignes; et inversement, une ligne peut contenir plusieurs titres. Finalement, les titres de pièces permettront d'enrichir notre modèle en vue de transcrire la totalité des informations disponibles au sein de ces documents.

---

[1] Brenner,The Théâtre Italien : its reportory, 1716-1793, University of California, Press, 1961.
[2] COÜASNON, Bertrand. DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way. *International Journal of Document Analysis and Recognition (IJDAR)*, 2006, vol. 8, no 2-3, p. 111-122.
[3] GRAVES, Alex, FERNÁNDEZ, Santiago, GOMEZ, Faustino, *et al.* Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In : *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006. p. 369-376.
[4] Arvanitopoulos, N., & Süsstrunk, S. (2014, September). Seam carving for text line extraction on color and grayscale historical manuscripts. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on (pp. 726-731). IEEE.
[5] A. Fischer, A. Keller, V. Frinken, and H. Bunke: "Lexicon-Free Handwritten Word Spotting Using Character HMMs," in Pattern Recognition Letters, Volume 33(7), pages 934-942, 2012.

[6] V. Romero, A. Fornés, N. Serrano, J.A. Sánchez, A.H. Toselli, V. Frinken, E. Vidal, J. Lladós. “The ESPOSALLES Database: An Ancient Marriage License Corpus for Off-line Handwriting Recognition”, Pattern Recognition, Volume 46, Issue 6, Pages 1658–1669, 2013.
[7] E. Augustin, J.-m. Brodin, M. Carr, E. Geoffrois, E. Grosicki, and F. Prłteux, “RIMES evaluation campaign for handwritten mail processing,” in Workshop on ICFHR, no. 1, 2006.