Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances

Hugo Scheithauer, Alix Chagué, Aurélia Rostaing, Lucas Terriel, Laurent Romary, Marie-Françoise Limon-Bonnet, Benjamin Davy, Gaetano Piraino, Franck Beltrami, Danis Habib, et al.

Les Futurs Fantastiques
Décembre 8-10, 2021
Bibliothèque nationale de France

Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances

Hugo Scheithauer, Alix Chagué, Aurélia Rostaing, Lucas Terriel, Laurent Romary, Marie-Françoise Limon-Bonnet, Benjamin Davy, Gaetano Piraino, Franck Beltrami, Danis Habib, Nathalie Denis et Marc Durand

Cet atelier proposera aux participant-es de prendre part à la production d'un modèle affiné de reconnaissance d'écriture manuscrite (REM) à partir de l'application eScriptorium/Kraken et de découvrir une méthodologie pour l'évaluation des performances des modèles de transcription. Un modèle affiné résulte du ré-entraînement d'un premier modèle générique à partir d'un autre jeu de données, avec comme objectif de le spécialiser dans un domaine particulier.

Après avoir chargé des numérisations de répertoires de notaires du Minutier central, nous appliquerons le modèle de transcription générique créé dans le cadre du projet LECTAUREP (Archives nationales - Inria). Ce modèle a été entraîné sur des documents administratifs écrits en français, datant des XIXe et XXe siècles, avec des mains différentes. Il s'agira ici de mesurer la performance du modèle sur des données issues d'un domaine qui lui est connu. Après ces premières observations, nous évaluerons la robustesse du modèle générique cette fois sur un jeu de données hors domaine, mais toujours en français et d'une période similaire. Nous utiliserons la librairie Python KaMI (Alix Chagué, Lucas Terriel) pour quantifier, entre autres, le taux d'erreurs au niveau des caractères (CER - *Character Error Rate*) et des mots (WER - *Word Error Rate*).

Après avoir produit un nouveau jeu de données de vérité de terrain adapté à notre échantillon, un modèle affiné sera entraîné et nous comparerons son efficacité selon la même méthodologie.

Il sera enfin intéressant d'étudier les différents types d'erreurs résiduelles produites par le modèle affiné, pour déterminer des stratégies de correction afin de se rapprocher au maximum de la vérité de terrain.

Cet atelier a donc trois objectifs, à savoir montrer le processus d'affinage d'un modèle de transcription dans eScriptorium ; montrer comment évaluer sa performance ; et enfin présenter de possibles post-traitements sur les sorties produites par la REM. Il s'adresse à quiconque est intéressé par l'application de la transcription automatique à des données patrimoniales. Le format encourageant l'échange, toute question sera la bienvenue.

https://gitlab.inria.fr/dh-projects/kami/kami-lib

Mots-clés : reconnaissance d'écriture manuscrite ; modèle affiné ; données patrimoniales ; post-traitements ; eScriptorium ; Kraken

******

Fantastic Futures
December 8-10, 2021
Bibliothèque nationale de France

Fine-tuning a handwritten text recognition model with eScriptorium and evaluating its performances

Hugo Scheithauer, Alix Chagué, Aurélia Rostaing, Lucas Terriel, Laurent Romary, Marie-Françoise Limon-Bonnet, Benjamin Davy, Gaetano Piraino, Franck Beltrami, Danis Habib, Nathalie Denis and Marc Durand

For this workshop, participants will take part in the fine-tuning of a handwritten text recognition (HTR) model with eScriptorium. Fine-tuning a model means retraining an initial generic model with a new dataset in order to specialize it in a particular domain.

After loading digitized notary records from the Minutier Central, we will apply the generic model created for the LECTAUREP project (Archives nationales - Inria). This model was trained on administrative documents written in French by different hands and dating from the 19th and 20th centuries. We will measure its performance on in-domain data. After this first experiment, the robustness of the generic model will be evaluated on an out-of-domain dataset made of handwritten documents still written in French and dating from a similar period. We will use a Python library called KaMI (Alix Chagué and Lucas Terriel) to measure the character error rate (CER) and the word error rate (WER).

After creating a new set of ground truth based on our out-of-domain sample, we will fine-tune the generic model and compare its efficiency using the same methodology.

Finally, and knowing that HTR models cannot achieve a perfect performance, we will observe the different types of residual errors made by the fine-tuned model, to then consider possible post-HTR corrections.

This workshop then has three objectives: to show how to fine-tune a transcription model with eScriptorium; to show how to evaluate its performance; and finally to present post-HTR possible corrections. It is intended for anyone who is interested in the application of

automatic transcription to patrimonial documents, and to enable the participants to exchange on the matter.

https://gitlab.inria.fr/dh-projects/kami/kami-lib

Keywords : handwritten text recognition ; fine-tuned model ; patrimonial documents ; post-HTR ; eScritptorium ; Kraken