Rédaction thèse : REN sur des romans français du xixème siècle

Caroline Koudoro-Parfait (1,2,3)

Sorbonne Université

26 septembre 2024

caroline.parfait@sorbonne-universite.fr,
(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France

Sommaire

✔ Introduction

0. Besoins et usages de la REN, le cas des utilisateurs des communautés littéraire et des SHS.
   - 0.1 Enquête Quantitative (Retour critique sur les erreurs)
   - 0.2 Enquête Qualitative
   - 0.3 Workshop

✔ I Analyse automatique de textes littéraires : enjeux et verrous de la REN

1. Circonscrire le périmètre de la tâche de REN spatiale sur un corpus Littéraire
   - 1.1 Épistémologie de la tâche de REN
     - 1.1.1 L'extraction d'information : comment on range le monde ?
     - 1.1.2 Catégories pour l'annotation : un lieu, des lieux
     - 1.1.3 Des systèmes en quête de robustesse
     - 1.1.4 Entité nommée de la linguistique au TAL : indéfinition
   - 1.2 Á la croisée du TAL et de la littérature : la REN spatiale une perspective HN
     - 1.2.1 Définir un nom de lieu en littérature
     - 1.2.2 comment catégoriser des lieux en littérature ?
     - 1.2.3 Des textes aux cartes : des géographies littéraires

✔ I Analyse automatique de textes littéraires : enjeux et verrous de la REN

✔ II Dépasser le problème du bruit de l'OCR en amont

4. Le bruit dans les sorties de REN : qualité du texte vs. qualité des modèles, les torts sont partagés
   - 4.1 Analyse des résultats de REN sur les textes de Référence : retours sur le silver standard
     - 4.1.1 Typologie des problèmes des modèles de REN
     - 4.1.2 Archéologie des systèmes de REN pour une évaluation plus fine des modèles
   - 4.2 Évaluation des systèmes de REN dans des conditions de labotatoire
     - 4.2.1 Un Gold pour ElteC, annotation en profondeur ou en largeur
     - 4.2.2 Un guide d'annotation pour les données contaminées
   - 4.3 Évaluation supervisées des résultats de la REN sur des données non contaminées par le bruit de l'OCR : de meilleurs Résultats ?
     - 4.3.1 Évaluation stricte ou souple
     - 4.3.2 Et pourtant le texte est propre le cas des FP
     - 4.3.3 Problèmes de modèles : Mauvais étiquetage des EN

✔ II Dépasser le problème du bruit de l'OCR en amont

5. Contourner le problème du bruit de l'OCR sur la REN en aval
   - 5.1 Des pistes pour la désambiguïsation automatique des formes contaminées des entités nommées spatiales
     - 5.1.1 Désambiguïsation de quoi s'agit-il ?
     - 5.1.2 Désambiguisation des formes contaminées des EN à l'aide de métrique de similarité, en route vers le clustering
     - 5.1.3 Entity Linking
   - 5.2 Epiméthée
     - 5.2.1 Outils d'Aide à la prise de décision (JE : Robustesse des systèmes + DH2023).Cluster
     - 5.2.2 Épiméthée : Une chaîne de traitement de bout-en-bout
     - 5.2.3 Des cas pratiques au Livrable

Rétro planning

	fin	chapitres
✔	29/10	0 Besoins et usages de la REN.
✔	**Relecture**	I.1 Circonscrire le périmètre de la tâche de REN
✔	**Relecture**	I.2 L'impact des erreurs d'OCR sur la REN
✔	21/10	II.3 Impact de la correction de l'OCR sur la REN
✔	23/10	II.4 Le bruit dans les sorties de REN en condition de labo.
✔	27/10	II.5 Contourner le problème du bruit de l'OCR sur la REN en aval
	31/10/2024	Envoi du manuscrit

Table – ✔ en cours de correction, ✔ rédaction avancée, ✔ squelette mis en place

Proposition de noms pour le jury

Prénom, Nom	poste
Andrea Del Lungo	PR Université Sorbonne
Ioana Galleron	PR Université Sorbonne Nouvelle - Paris 3
Elena Pierazzo	PR Université CESR Tours
Carmen Brando	
Antoine Doucet rapporteur	
...	...