Mapping spatial named entities from noisy OCR output: Epiméthée from OCR to map.

Caroline Koudoro-Parfait (1,2,3), Motasem Alrahabi (1), Yoann Dupont(4), Gaël Lejeune (2), Glenn Roe (1)

DH2023, Collaboration as opportunity

July 2023

caroline.parfait@sorbonne-universite.fr, gael.lejeune@sorbonne-universite.fr, yoa.dupont@gmail.com, motasem.alrahabi@sorbonne-universite.fr, glenn.roe@sorbonne-universite.fr
(1) OBTIC, Sorbonne Université, Paris, France
(2) STIH, Sorbonne Université, Paris France
(3) SCAI, Sorbonne Center for Artificial Intelligence, Paris, France
(4) Lattice, 1 Rue Maurice Arnoux, 92120 Montrouges, France

Can NER perform well on noisy OCR transcripts ?

Question How to assess the quality of NER results on noisy OCR corpora ? [Koudoro-Parfait et al., 2021]

- Data : Eltec reference texts and OCR versions
- Manual evaluation : all misspellings are not of equal importance
- Intersections reference/OCR ➜ **contaminated NE issues**

Can NER perform well on noisy OCR transcripts ?

Question How to assess the quality of NER results on noisy OCR corpora ? [Koudoro-Parfait et al., 2021]

- Data : Eltec reference texts and OCR versions
- Manual evaluation : all misspellings are not of equal importance
- Intersections reference/OCR ➜ **contaminated NE issues**

A search for appropriate solutions to make NER results on noisy data usable :

Correcting OCR transcriptions to improve NER quality ✗

Linking contaminated entities to knowledge bases entities ✗

Charting entities get an overview ✓ [Koudoro-Parfait and Lejeune, 2024]

Combining the results of **several NER models** ✓

Disambiguating using similarity metrics and clusters ✓ [Koudoro-Parfait et al., 2022], [Koudoro-Parfait, 2022]

Charting NE from low quality OCR (CER : 0.14)

(a) NER on reference version

(b) manual filtering on reference

(c) NER on Kraken version

(d) manual filtering on Kraken

Figure – *"La petite Jeanne"*, Carraud, 1853. Dark green / red : 3 NER tool, Middle green /orange : 2 NER tools, green-grey / yellow : 1 NER tool.

Charting NE from high quality OCR (CER : 0.05)

(a) NER on reference version

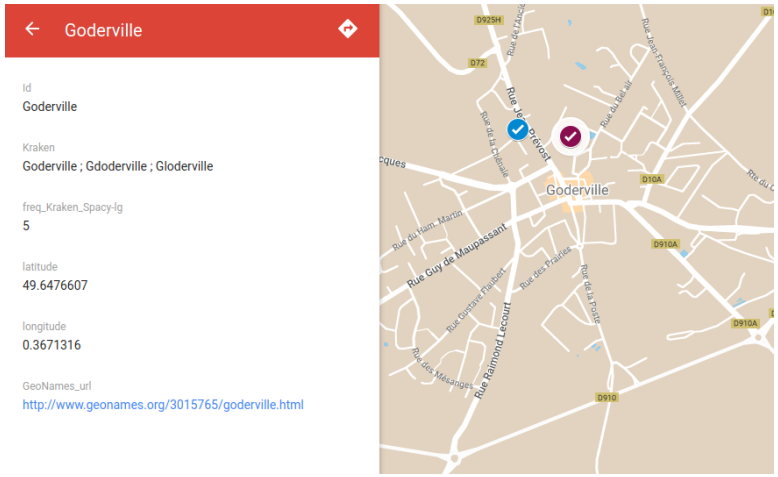
(b) manual filtering on reference

(c) NER on Kraken version

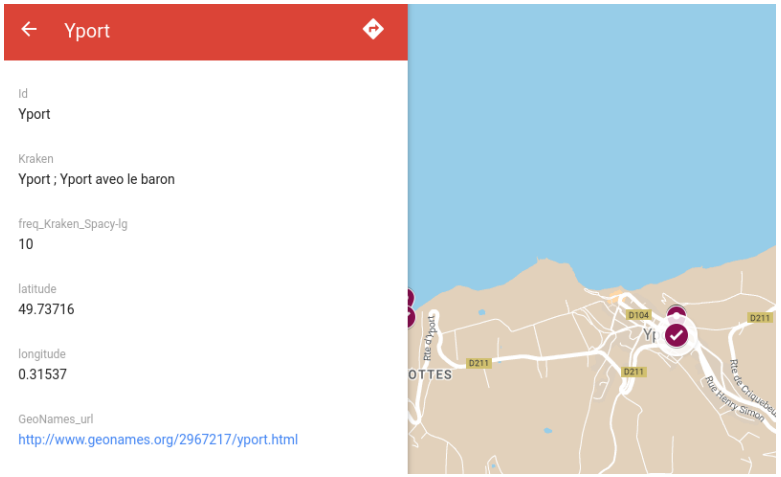
(d) manual filtering on Kraken

Figure – *"Le petit chose"*, Daudet,1868. Dark green / red : 3 NER tool, Middle green /orange : 2 NER tools, green-grey / yellow : 1 NER tool.

Aligning contaminated forms of named entities (NE)



(a) Contaminated forms of `Goderville`



(b) Contaminated forms of `Yport`

- Average text quality (CER : 0.12)
- Most forms of Frequent NE are extracted (Jaccard : 0.8, Cosine : 0.45)

"Une vie", Maupassant, 1883.

Clustering to help manual filtering

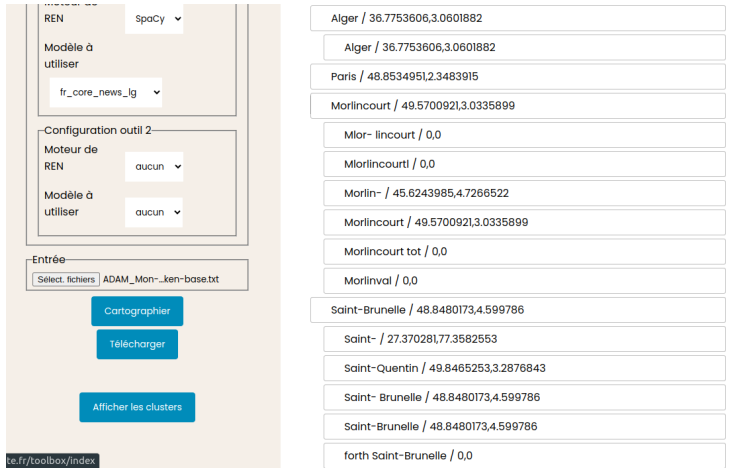
Version	Centroid	Cluster members
Réf.[1]	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken	Montparnasse	Montparnasse, boulevard Montparnasse, theatre Montparnasse, Gymnase, Debarrassez, WWt3, rs5, ytP
Tess fr	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken[1]	quartier Latin	Quartier Latin, quartier Latin, Quartier, ILatinl, Latin, Cachotier, Moustier, Orties, PARTIE=.
Tess fr	quartier Latin	Quartier Latin, quartier Latin, Quartier, Latin, Latin !., Moustier, Orties,
Kraken[2]	Goderville	Gdoderville, Gloderville, Goderville, Barville, Fourville, OD0

Table – Groups where centroid is a TP

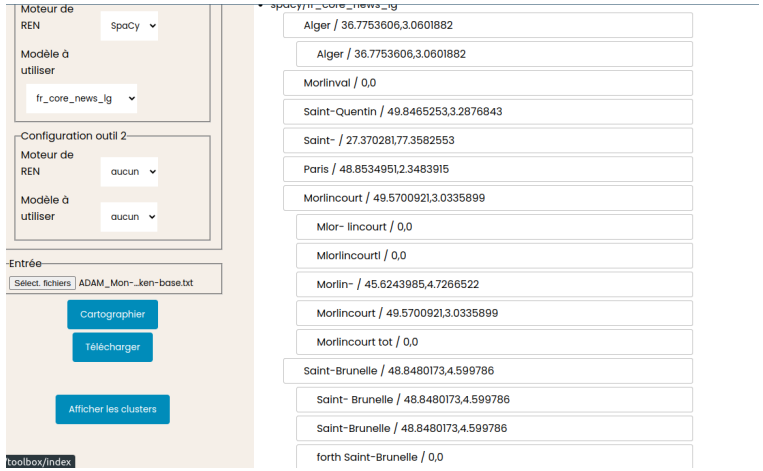
- These are the best results.
- Some results have as centroid a FP or an ambiguous entity.

[1] `spaCy_lg`, *"Le petit chose"*, Daudet, 1868. [2] *`spaCy_lg`*, *"Une vie"*, G. de Maupassant, 1883

Using maps to improve clusters and filter out FP



(a) Before user's action



(b) After user's action

Contributions

- Épiméthée an end-to-end workflow
  - OCR → NER → Map
  - Clustering
  - Interactive filtering assistance tools
  - Reusable outputs

    → Humans can deal with noise, but what if there is silence ?
- Clustering : helps decision making for **linking contaminated NE**
  - Manual evaluation
  - Observation of **borderline cases** : Centroid is FP or Ambiguous
  - Automated cluster improvement : **filtering out outliers**

Contributions

- Épiméthée an end-to-end workflow
  - OCR → NER → Map
  - Clustering
  - Interactive filtering assistance tools
  - Reusable outputs

    → Humans can deal with noise, but what if there is silence ?
- Clustering : helps decision making for **linking contaminated NE**
  - Manual evaluation
  - Observation of **borderline cases** : Centroid is FP or Ambiguous
  - Automated cluster improvement : **filtering out outliers**

Perspectives :

➥ Improving automated geolocalization.

➥ Evaluation of tools in other languages

→ Difficulties in providing access to tools for Low-resource language

#additional# Clustering to help manual filtering

Version	Centroid	Cluster members
Réf.[1]	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken	Montparnasse	Montparnasse, boulevard Montparnasse, theatre Montparnasse, Gymnase, Debarrassez, WWt3, rs5, ytP
Tess fr	Montparnasse	Montparnasse, boulevard Montparnasse, théâtre Montparnasse, Montmartre, rue Bonaparte, Mont-, Saumon, Gymnase
Kraken[1]	quartier Latin	Quartier Latin, quartier Latin, Quartier, lLatinl, Latin, Cachotier, Moustier, Orties, PARTIE=.
Tess fr	quartier Latin	Quartier Latin, quartier Latin, Quartier, Latin, Latin !., Moustier, Orties,
Kraken[2]	Goderville	Gdoderville, Gloderville, Goderville, Barville, Fourville, OD0

(a) Groups where centroid is a TP

Version	Centroid	Cluster members
Réf.[1]	PION	Lyon, Odéon, Rio, PION
Kraken	Fougeroux	Luxembourg, Perou, Broum, Fougeroux, MY, Vaudoux, les Fougeroux
Tess fr	Viens	Vienne, Viens, Agen, ENLÈVEMENT, Fe, Providence, Tiens

(b) Groups where centroid is a FP

- These are the best results.
- Some results have as centroid a FP or an ambiguous entity.

[1] spaCy lg *"Le petit chose"* Daudet 1868 [2] *spaCy lg "Une vie"* G. de Maupassant 1883

References I

Koudoro-Parfait, C. (2022).
Évaluation de la tâche de clustering pour l'alignement de formes contaminées d'entités nommées issues d'un corpus ocr bruité.
Actes de la journée d'étude sur la robustesse des systemes de TAL, page 5.

Koudoro-Parfait, C. and Lejeune, G. (2024).
Reconnaissance des Entités Nommées spatiales sur un corpus littéraire bruité : des entités à la carte.
In *In Proceedings of the Séminaire des sources aux Systèmes d'Information Géographique*.

References II

Koudoro-Parfait, C., Lejeune, G., and Buth, R. (2022).
Reconnaissance d'entités nommées sur des sorties ocr bruitées : des pistes pour la désambiguïsation morphologique automatique (resolution of entity linking issues on noisy ocr output : automatic disambiguation tracks).
In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 45–55.

Koudoro-Parfait, C., Lejeune, G., and Roe, G. (2021).
Spatial named entity recognition in literary texts : What is the influence of OCR noise ?
In Moncla, L., Brando, C., and McDonough, K., editors, *GeoHumanities@SIGSPATIAL 2021 : Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021*, pages 13–21. ACM.