OCR impact on Named Entity Recognition on historical documents : Problem Presentation ... and Epistemological consequences

Caroline Koudoro-Parfait (1,2,3) &, Gaël Lejeune (2, 4)

Summer School on Digital Humanities (UNAM/SCAI)

September 13 2023

caroline.parfait,gael.lejeune@sorbonne-universite.fr

(1) OBTIC, Observatory of Texts and Corpus ;
(2) STIH, Sense, Text, Informatics and History
(3) SCAI, Sorbonne Center for Artificial Intelligence ;
(4) CERES, Center of Research on digital methods for Social Sciences

Outline

Finding places and persons in literary texts

Challenge How we able to process large (multilingual) literary corpora in order to find information about persons and locations ?

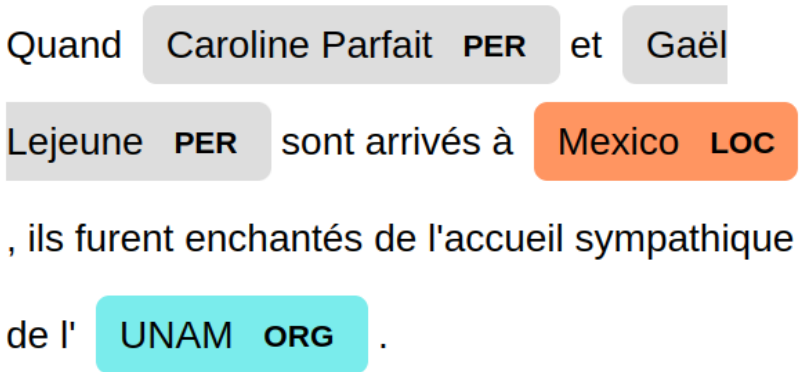
Finding places and persons in literary texts

Challenge How we able to process large (multilingual) literary corpora in order to find information about persons and locations ?

Step 1 OCR : Optical Character Recognition (images → plain text)

Step 2 NER : Named Entity Recognition (plain text → annotated text)

Quand Caroline Parfait et Gaël Lejeune sont arrivés à Mexico, ils furent enchantés de l'accueil sympathique de l'UNAM.



Finding places and persons in literary texts

Challenge How we able to process large (multilingual) literary corpora in order to find information about persons and locations ?

Step 1 OCR : Optical Character Recognition (images → plain text)

Step 2 NER : Named Entity Recognition (plain text → annotated text)

Quand Caroline Parfait et Gaël Lejeune sont arrivés à Mexico, ils furent enchantés de l'accueil sympathique de l'UNAM.

Quand Caroline Parfait PER et Gaël Lejeune PER sont arrivés à Mexico LOC , ils furent enchantés de l'accueil sympathique de l' UNAM ORG .

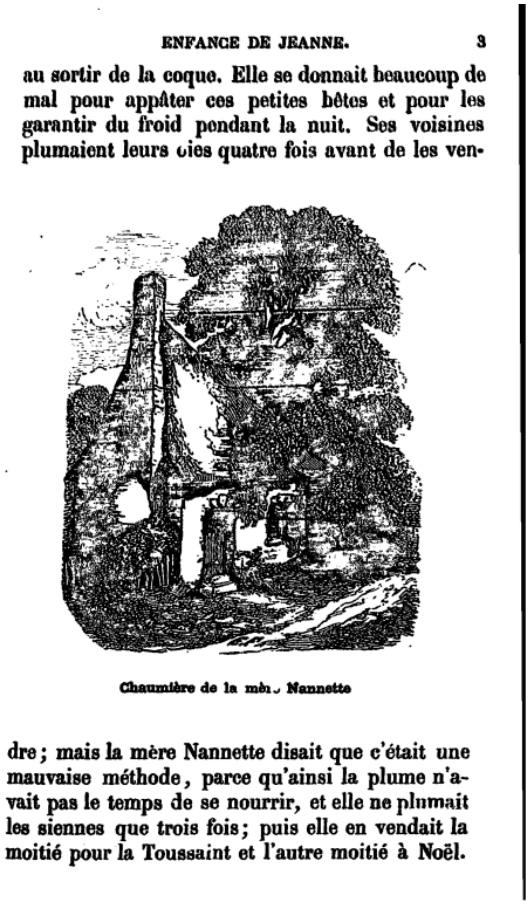
➜ In theory, we just have to use OCR to transform a Computer Vision problem into a solved (?) NLP problem

OCR : Theory and Practice

In theory there is no difference between theory and practice but in practice there is. Yogi Berra

OCR : Theory and Practice

In theory there is no difference between theory and practice but in practice there is. Yogi Berra



ENFANCE DE JEANNE. 3

au sortir de la coque. Elle se donnait beaucoup de mal pour appâter ces petites bêtes et pour les garantir du froid pendant la nuit. Ses voisines plumaient leurs oies quatre fois avant de les ven-

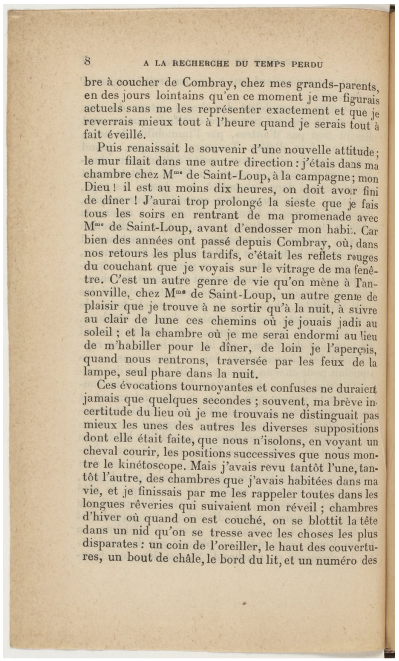
Chaumière de la mère Nannette

dre ; mais la mère Nannette disait que c'était une mauvaise méthode, parce qu'ainsi la plume n'avait pas le temps de se nourrir, et elle ne plumait les siennes que trois fois ; puis elle en vendait la moitié pour la Toussaint et l'autre moitié à Noël.

Kraken (OCR)
Ses voisines plumaient leurs oioes quatre fois avant de les **ven-** LL I II II I I IIF M ii I I II E E g Chamnnlhrs de t a mn Mamnnetta **dre**; mais la mere Nannette disait que eetait une mauvaise m6thode, paree qu'ainsi la plume ...

Green : illustration, blue : caption of the illustration. Carraud, French ELTeC corpus.

NER : Theory and Practice



8 A LA RECHERCHE DU TEMPS PERDU

bre à coucher de Combray, chez mes grands-parents, en des jours lointains qu'en ce moment je me figurais actuels sans me les représenter exactement et que je reverrais mieux tout à l'heure quand je serais tout à fait éveillé.

Puis renaissait le souvenir d'une nouvelle attitude; le mur filait dans une autre direction : j'étais dans ma chambre chez Mme de Saint-Loup, à la campagne; mon Dieu ! il est au moins dix heures, on doit avoir fini de dîner ! J'aurai trop prolongé la sieste que je fais tous les soirs en rentrant de ma promenade avec Mme de Saint-Loup, avant d'endosser mon habi:. Car bien des années ont passé depuis Combray, où, dans nos retours les plus tardifs, c'était les reflets rouges du couchant que je voyais sur le vitrage de ma fenêtre. C'est un autre genre de vie qu'on mène à Tansonville, chez Mme de Saint-Loup, un autre genre de plaisir que je trouve à ne sortir qu'à la nuit, à suivre au clair de lune ces chemins où je jouais jadis au soleil ; et la chambre où je me serai endormi au lieu de m'habiller pour le dîner, de loin je l'aperçois, quand nous rentrons, traversée par les feux de la lampe, seul phare dans la nuit.

Ces évocations tournoyantes et confuses ne duraient jamais que quelques secondes ; souvent, ma brève incertitude du lieu où je me trouvais ne distinguait pas mieux les unes des autres les diverses suppositions dont elle était faite, que nous n'isolons, en voyant un cheval courir, les positions successives que nous montre le kinétoscope. Mais j'avais revu tantôt l'une, tantôt l'autre, des chambres que j'avais habitées dans ma vie, et je finissais par me les rappeler toutes dans les longues rêveries qui suivaient mon réveil ; chambres d'hiver où quand on est couché, on se blottit la tête dans un nid qu'on se tresse avec les choses les plus disparates : un coin de l'oreiller, le haut des couvertures, un bout de châle, le bord du lit, et un numéro des

Source gallica.bnf.fr / Bibliothèque nationale de France

Figure: A page from M. Proust

Puis renaissait le souvenir d'une nouvelle attitude ;

le mur filait dans une autre direction : j'étais dans ma

chambre chez M PER ** de Saint-Loup LOC , ala campagne; mon

Dieu! MISC il est au moins dix heures, on doit avoir fini

de diner ! J'aurai trop prolongé la sieste que je fais

tous les soirs en rentrant de ma promenade avec

M LOC ™ de Saint-Loup LOC , avant d'endosser mon habi:. Car

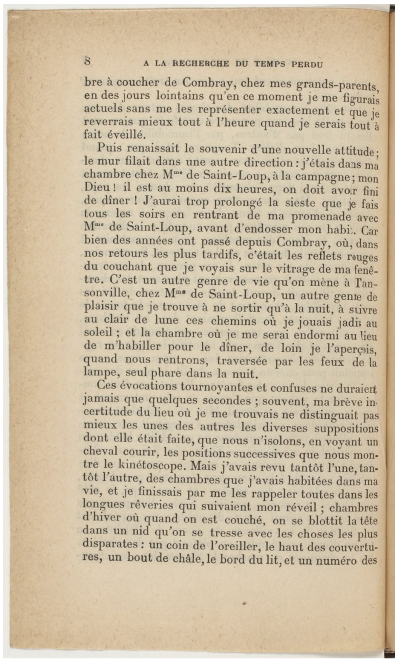
bien des années ont passé depuis Combray LOC , ot, dans

nos retours les plus tardifs, c'était les reflets rouges

du couchant que je voyais sur le vitrage de ma fené-

Figure: Extract of NER output

NER : Theory and Practice



8 A LA RECHERCHE DU TEMPS PERDU

bre à coucher de Combray, chez mes grands-parents, en des jours lointains qu'en ce moment je me figurais actuels sans me les représenter exactement et que je reverrais mieux tout à l'heure quand je serais tout à fait éveillé.

Puis renaissait le souvenir d'une nouvelle attitude; le mur filait dans une autre direction : j'étais dans ma chambre chez Mme de Saint-Loup, à la campagne; mon Dieu ! il est au moins dix heures, on doit avoir fini de dîner ! J'aurai trop prolongé la sieste que je fais tous les soirs en rentrant de ma promenade avec Mme de Saint-Loup, avant d'endosser mon habi:. Car bien des années ont passé depuis Combray, où, dans nos retours les plus tardifs, c'était les reflets rouges du couchant que je voyais sur le vitrage de ma fenêtre. C'est un autre genre de vie qu'on mène à Tansonville, chez Mme de Saint-Loup, un autre genre de plaisir que je trouve à ne sortir qu'à la nuit, à suivre au clair de lune ces chemins où je jouais jadis au soleil ; et la chambre où je me serai endormi au lieu de m'habiller pour le dîner, de loin je l'aperçois, quand nous rentrons, traversée par les feux de la lampe, seul phare dans la nuit.

Ces évocations tournoyantes et confuses ne duraient jamais que quelques secondes ; souvent, ma brève incertitude du lieu où je me trouvais ne distinguait pas mieux les unes des autres les diverses suppositions dont elle était faite, que nous n'isolons, en voyant un cheval courir, les positions successives que nous montre le kinétoscope. Mais j'avais revu tantôt l'une, tantôt l'autre, des chambres que j'avais habitées dans ma vie, et je finissais par me les rappeler toutes dans les longues rêveries qui suivaient mon réveil ; chambres d'hiver où quand on est couché, on se blottit la tête dans un nid qu'on se tresse avec les choses les plus disparates : un coin de l'oreiller, le haut des couvertures, un bout de châle, le bord du lit, et un numéro des

Source gallica.bnf.fr / Bibliothèque nationale de France

Figure: A page from M. Proust

Puis renaissait le souvenir d'une nouvelle attitude ;

le mur filait dans une autre direction : j'étais dans ma

chambre chez M PER ** de Saint-Loup LOC , ala campagne; mon

Dieu! MISC il est au moins dix heures, on doit avoir fini

de diner ! J'aurai trop prolongé la sieste que je fais

tous les soirs en rentrant de ma promenade avec

M LOC ™ de Saint-Loup LOC , avant d'endosser mon habi:. Car

bien des années ont passé depuis Combray LOC , ot, dans

nos retours les plus tardifs, c'était les reflets rouges

du couchant que je voyais sur le vitrage de ma fené-

Figure: Extract of NER output

➜ Not so bad uh ?

NER : Theory and Practice

Next paragraph : not so good

tre. C'est un autre genre de vie qu'on méne & Tan-

sonville LOC , chez M LOC ™* de Saint-Loup LOC , un autre gente de

plaisir que je trouve A LOC ne sortir qu'a la nuit, a stivre

au clair de lune ces chemins ot je jouais jadis au

soleil ; et la chambre ot je me serai endormi au lieu

de m'habiller pour le diner, de loin Je Vapereis LOC ,

quand nous rentrons, traversée par les feux de la

lampe, seul phare dans la nuit.

Figure: A page from M. Proust

NER : Theory and Practice

Next paragraph : not so good

tre. C'est un autre genre de vie qu'on méne & Tan-

sonville LOC , chez M LOC ™* de Saint-Loup LOC , un autre gente de

plaisir que je trouve A LOC ne sortir qu'a la nuit, a stivre

au clair de lune ces chemins ot je jouais jadis au

soleil ; et la chambre ot je me serai endormi au lieu

de m'habiller pour le diner, de loin Je Vapereis LOC ,

quand nous rentrons, traversée par les feux de la

lampe, seul phare dans la nuit.

Figure: A page from M. Proust

- Hyphenation
- Tokenization
- OOV enhances noise
- True casing would be needed

NER : Theory and Practice

Next paragraph : not so good

tre. C'est un autre genre de vie qu'on méne & Tan-

sonville LOC , chez M LOC ™* de Saint-Loup LOC , un autre gente de

plaisir que je trouve A LOC ne sortir qu'a la nuit, a stivre

au clair de lune ces chemins ot je jouais jadis au

soleil ; et la chambre ot je me serai endormi au lieu

de m'habiller pour le diner, de loin Je Vapereis LOC ,

quand nous rentrons, traversée par les feux de la

lampe, seul phare dans la nuit.

Figure: A page from M. Proust

- Hyphenation
- Tokenization
- OOV enhances noise
- True casing would be needed

➜ HOW much guilty OCR is ?

- Kraken (default)
- Tesseract (default)
- Tesseract French, Portuguese and English

What we observe

➜ when checking NER outputs – spaCy-lg [1] and stanza[2] – on noisy OCR

Impact types	**contexte**	**spaCy_lg**	`stanza`
Orthographic Contamination inside the entity	*il en est tombe au sort cinq de Sainl-Bruncl le* *duranta todo o tompe em qne ostivesso em Portngal*	Sainl-Bruncl le. Portngal	Sainl-Bruncl le N/A
adding a lowercase character	*Aux kEtats-Unis*	()	kEtats-Unis
punctuation substituted by a character	*about Manchesterl A pretty state*	Manchesterl	Manchesterl
Truncated term	*dans l'intérieur de l'Améri-* *et le golfe de Cali-foruie._n*	Améri— golfe de Cali-	() golfe de Cali-
concatenated words	*[...] larue Saint-Honoré;* *afriver aMorlincourt' tot*	_ Saint-Honoré ()	() ()

Table: Non-exhaustive list of OCR errors that can have impact on NER. Corpora : ELTeC French, Portuguese and English[3]

[1] `https://spacy.io/usage/linguistic-features`
[2] `https://stanfordnlp.github.io/stanza/ner.html`
[3] `https://zenodo.org/communities/eltec/`

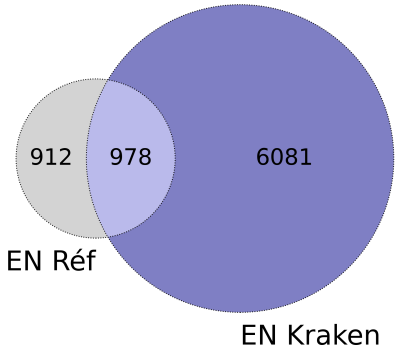
What we observe

concerning contaminated entities : Guadelajara (Ref.), Guadelazara (Kraken), Guadelaxara (Tess.) and Guadelaaew*ra (Tess.fr)

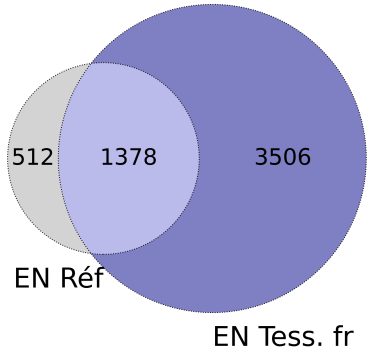
	REF		Kraken		Tess. fr	
	`stanza`	`spacy_lg`	`stanza`	`spacy_lg`	`stanza`	`spacy_lg`
Nb. types	314	337	654	816	496	393
VP	156	156	**190**	**177**	**190**	153
FP	158	181	464	639	306	240

Table: Annotation of True Positives and False Positives on named entity sets for the reference and OCR versions. Daudet, ELTeC fr.

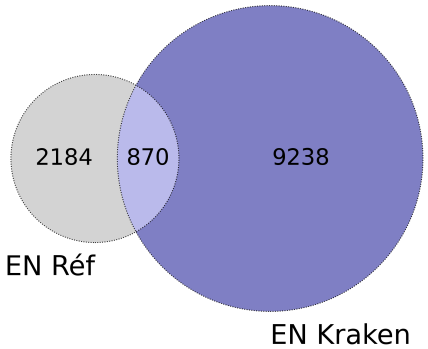
What we observe



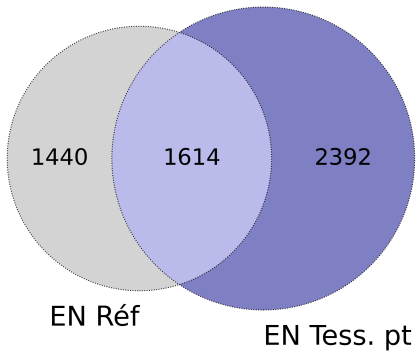
(a) ELTeC french Kraken–spaCy-lg



(b) ELTeC french Tess. fr–spacy-lg



(c) ELTeC Portuguese Kraken–spaCy-lg



(d) ELTeC Portuguese Tess. pt–spacy-lg

Figure: intersections: evaluation of the NER outputs with spaCy-lg for Kraken and Tesseract fr. on ELTeC French and Portuguese

What we observe

➜ when we automatically correct the OCR

	#Entités		Évaluation par NERVAL			
Version	OCR	Réf.	Intersection	Précision	Rappel	F_1 mesure
Kraken	1122	944	566	0.504	0.761	*0.607*
Tess fr	860	944	646	0.751	0.868	**0.805**
Tess	920	944	597	0.649	0.802	*0.718*
Kraken + Jspllfr	1027	944	471	0.459	0.633	0.532 ⇓
Tess fr + Jspllfr	794	944	532	0.67	0.715	**0.692** ⇓
Tess + Jspllfr	846	944	503	0.595	0.676	*0.633* ⇓

Table: Nerval[4] a tool for evaluating NER on noisy data. Daudet, ELTeC Fr.

[4] `https://gitlab.com/teklia/ner/nerval`

NLP pipeline : what is happening ?

Hypothetical exchange between CS researcher and DH researcher

NLP pipeline : what is happening ?

Hypothetical exchange between CS researcher and DH researcher
Remember Gemma: *Outputs are not directly proportional to the input*

NLP pipeline : what is happening ?

Hypothetical exchange between CS researcher and DH researcher

Remember Gemma: *Outputs are not directly proportional to the input*

Think about robustness !

- DH: The results are disappointing !
- CS: This is not my fault, your data is so creepy !

Think about robustness !

- DH: The results are disappointing !
- CS: This is not my fault, your data is so creepy !
- DH : Adapt your approach to data [Koudoro-Parfait, 2022]

Think about robustness !

- DH: The results are disappointing !
- CS: This is not my fault, your data is so creepy !
- DH : Adapt your approach to data [Koudoro-Parfait, 2022]

Remember Gerardo *What do you need in your corpus ?*

Think about robustness !

- DH: The results are disappointing !
- CS: This is not my fault, your data is so creepy !
- DH : Adapt your approach to data [Koudoro-Parfait, 2022]

Remember Gerardo *What do you need in your corpus ?*

Observing OCR noise

Reference

bre **à** coucher de Combray, chez mes grands-parents, en des jours lointains qu'en ce moment je me figurais actuels sans me les représenter exactement et que je reverrais mieux tout a l'heure quand je serais tout **à** fait éveillé.
Puis renaissait le souvenir d'une nouvelle attitude ; le mur filait dans une autre direction : j'étais dans ma chambre chez Mme de Saint-Loup, **à la** campagne; mon Dieu! il est au moins dix heures, on doit avoir fini de dîner ! J'aurai trop prolongé la sieste que je fais tous les soirs en rentrant de ma promenade avec **Mme** de Saint-Loup, avant d'endosser mon **habit**.
Car bien des années ont passé depuis Combray, où, dans nos retours les plus tardifs, c'était les reflets rouges du couchant que je voyais sur le vitrage de ma **fenê- tre**.
C'est un autre genre de vie qu'on mène **à** Tan- sonville, chez **Mme** de Saint-Loup, un autre genre de plaisir que je trouve à ne sortir qu'**à** la nuit, **à suivre** au clair de lune ces chemins **où** je jouais

Default Tesseract

8 A LA RECHERCHE DU TEMPS PERDU
bre **&** coucher de Combray, chez mes grands-parents, en des jours lointains qu'en ce moment je me figurais actuels sans me les représenter exactement et que je reverrais mieux tout a l'heure quand je serais tout **4** fait éveillé.
Puis renaissait le souvenir d'une nouvelle attitude ; le mur filait dans une autre direction : j'étais dans ma chambre chez **M**** de Saint-Loup, **ala** campagne; mon Dieu! il est au moins dix heures, on doit avoir fini de diner ! J'aurai trop prolongé la sieste que je fais tous les soirs en rentrant de ma promenade avec M™ de Saint-Loup, avant d'endosser mon **habi:**.
Car bien des années ont passé depuis Combray, **ot**, dans nos retours les plus tardifs, c'était les reflets rouges du couchant que je voyais sur le vitrage de ma **fené- tre**.
C'est un autre genre de vie qu'on **méne &** Tan- sonville, chez **M™*** de Saint-Loup, un autre **gente** de plaisir que je trouve **A** ne sortir qu'**a** la nuit, a **stivre** au clair de lune ces chemins **ot** je jouais

Tesseract_fr

8 A LA RECHERCHE DU TEMPS PERDU
bre **à** coucher de Combray, chez mes grands-parents, en des jours lointains qu'en ce moment je me figurais actuels sans me les représenter exactement et que je reverrais mieux tout à l'heure quand je serais tout **à** fait éveillé.
Puis renaissait le souvenir d'une nouvelle attitude ; le mur filait dans une autre direction : j'étais dans ma chambre chez **M°*** de Saint-Loup, à la campagne; mon Dieu ! il est au moins dix heures, on doit avoir fini de dîner ! J'aurai trop prolongé la sieste que je fais tous les soirs en rentrant de ma promenade avec **M°*** de Saint-Loup, avant d'endosser mon **habi:**.
Car bien des années ont passé depuis Combray, **où**, dans nos retours les plus tardifs, c'était les reflets rouges du couchant que je voyais sur le vitrage de ma **fené- tre**.
C'est un autre genre de vie qu'on **mène à l'an-** sonville, chez **Me** de Saint-Loup, un autre **genie** de plaisir que je trouve à ne sortir qu'à la nuit, **à suvre** au clair de lune ces chemins **où** je jouais

Examples of OCR noise

- à $\rightarrow$ & [twice]
- à $\rightarrow$ A
- à $\rightarrow$ a
- à $\rightarrow$ 4
- à la $\rightarrow$ ala

Examples of OCR noise

- à → & [twice]
- à → A
- à → a
- à → 4
- à la → ala
-
- mm → ™
- mm → **
- mm → ™*

Examples of OCR noise

- à → & [twice]
- à → A
- à → a
- à → 4
- à la → ala
-
- mm → ™
- mm → **
- mm → ™*
-
- t → : (habit → habi:)
- é → è
- ...

Examples of OCR noise

- à → & [twice]
- à → A
- à → a
- à → 4
- à la → ala
-
- mm → ™
- mm → **
- mm → ™*
-
- t → : (habit → habi:)
- é → è
- . . .
- aperçois → Vapereis
- l'autre → autre
- d'hiver → dhiver

How to measure OCR noise ?

Word Error rate and Character Error rate (supervised), Lexical rate (semi-supervised)

How to measure OCR noise ?

Word Error rate and Character Error rate (supervised), Lexical rate (semi-supervised)

Results for default Tesseract :

- WER : 0.1561
- CER : 0.0682

How to measure OCR noise ?

Word Error rate and Character Error rate (supervised), Lexical rate (semi-supervised)

Results for default Tesseract :

- WER : 0.1561
- CER : 0.0682
- LEX : 0.8318 (ref = 0.8971) with lowercase = True
- LEX : 0.7578 (ref = 0.8654) with lowercase = False

How to measure OCR noise ?

Word Error rate and Character Error rate (supervised), Lexical rate (semi-supervised)

Results for default Tesseract :

- WER : 0.1561
- CER : 0.0682
- LEX : 0.8318 (ref = 0.8971) with lowercase = True
- LEX : 0.7578 (ref = 0.8654) with lowercase = False

Results for Tesseract_fr :

- WER : 0.1369 (-0.02)
- CER : 0.0609 (-0.0073)
- LEX : 0.8645 (+0.03), LEX for ref = 0.8971 (lowercase = True)
- LEX : 0.8009 (+0.04), LEX for ref = 0.8654 (lowercase = False)

Visual observation of NER on different versions

bre à coucher de Combray LOC , chez mes grands-parents,
en des jours lointains qu'en ce moment je me figurais
actuels sans me les représenter exactement et que je
reverrais mieux tout a l'heure quand je serais tout à
fait éveillé.

Puis renaissait le souvenir d'une nouvelle attitude ;
le mur filait dans une autre direction : j'étais dans ma
chambre chez Mme de Saint-Loup PER , à la campagne; mon
Dieu! MISC il est au moins dix heures, on doit avoir fini
de dîner ! J'aurai trop prolongé la sieste que je fais
tous les soirs en rentrant de ma promenade avec
Mme de Saint-Loup PER , avant d'endosser mon habit. Car
bien des années ont passé depuis Combray LOC , où, dans
nos retours les plus tardifs, c'était les reflets rouges
du couchant que je voyais sur le vitrage de ma fenê-
tre. C'est un autre genre de vie qu'on mène à Tan-
sonville LOC , chez Mme de Saint-Loup PER , un autre gente de
plaisir que je trouve à ne sortir qu'a la nuit, à suivre
au clair de lune ces chemins où je jouais jadis au
soleil ; et la chambre où je me serai endormi au lieu
de m'habiller pour le dîner, de loin je l'aperçois,
quand nous rentrons, traversée par les feux de la
lampe, seul phare dans la nuit.

8 A LA RECHERCHE DU TEMPS MISC PERDU

bre \& coucher de Combray LOC , chez mes grands-parents,
en des jours lointains qu'en ce moment je me figurais
actuels sans me les représenter exactement et que je
reverrais mieux tout a l'heure quand je serais tout 4
fait éveillé.

Puis renaissait le souvenir d'une nouvelle attitude ;
le mur filait dans une autre direction : j'étais dans ma
chambre chez M PER ** de Saint-Loup LOC , ala campagne;
mon
Dieu! MISC il est au moins dix heures, on doit avoir fini
de diner ! J'aurai trop prolongé la sieste que je fais
tous les soirs en rentrant de ma promenade avec
M LOC ™ de Saint-Loup LOC , avant d'endosser mon habi:. Car
bien des années ont passé depuis Combray LOC , ot, dans
nos retours les plus tardifs, c'était les reflets rouges
du couchant que je voyais sur le vitrage de ma fené-
tre. C'est un autre genre de vie qu'on méne \& Tan-
sonville LOC , chez M LOC ™* de Saint-Loup LOC , un autre
gente de
plaisir que je trouve A LOC ne sortir qu'a la nuit, a stivre
au clair de lune ces chemins ot je jouais jadis au
soleil ; et la chambre ot je me serai endormi au lieu
de m'habiller pour le diner, de loin Je Vapereis LOC ,
quand nous rentrons, traversée par les feux de la
lampe, seul phare dans la nuit.

8 A LA RECHERCHE DU TEMPS MISC PERDU

bre à coucher de Combray LOC , chez mes grands-parents,
en des jours lointains qu'en ce moment je me figurais
actuels sans me les représenter exactement et que je
reverrais mieux tout à l'heure quand je serais tout à
fait éveillé.

Puis renaissait le souvenir d'une nouvelle attitude ;
le mur filait dans une autre direction : j'étais dans ma
chambre chez M°* de MISC Saint-Loup LOC , à la campagne;
mon
Dieu ! MISC il est au moins dix heures, on doit avoir fini
de dîner ! J'aurai trop prolongé la sieste que je fais
tous les soirs en rentrant de ma promenade avec
M LOC °* de Saint-Loup LOC , avant d'endosser mon habi:. Car
bien des années ont passé depuis Combray LOC , où, dans
nos retours les plus tardifs, c'était les reflets rouges
du couchant que je voyais sur le vitrage de ma fené-
tre. C'est un autre genre de vie qu'on mène à l'an-
sonville, chez Me de Saint-Loup PER , un autre genie de
plaisir que je trouve à ne sortir qu'à la nuit, à suvre
au clair de lune ces chemins où je jouais jadis au
soleil ; et la chambre où je me serai endormi au lieu
de m'habiller pour le dîner, de loin je l' aperçis LOC ,
quand nous rentrons, traversée par les feux de la
lampe, seul phare dans la nuit.

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

For NLP : There is no such thing as clean data !

→ Make it (your approach) Simple, Robust, **Multilingual**.

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

For NLP : There is no such thing as clean data !

→ Make it (your approach) Simple, Robust, **Multilingual**.

- Remember **Fernanda** : character n-grams are useful features

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

For NLP : There is no such thing as clean data !

→ Make it (your approach) Simple, Robust, **Multilingual**.

- Remember **Fernanda** : character n-grams are useful features
- Remember **Andric** : there are different ways to represent data

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

For NLP : There is no such thing as clean data !

→ Make it (your approach) Simple, Robust, **Multilingual**.

- Remember **Fernanda** : character n-grams are useful features
- Remember **Andric** : there are different ways to represent data
- → particularly useful outside laboratory conditions or tutorial data (so-called recipes) !

What do we need in applications ?

When we develop : Make it work, make it clean, make it fast

For NLP : There is no such thing as clean data !

→ Make it (your approach) Simple, Robust, **Multilingual**.

- Remember **Fernanda** : character n-grams are useful features
- Remember **Andric** : there are different ways to represent data
- → particularly useful outside laboratory conditions or tutorial data (so-called recipes) !

Examples outside OCR : user generated data (Twitter/X), Automated Speech Recognition data, multilingual corpus and other kinds of heterogenous data . . .

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What works

- NER can handle (reasonably) noisy OCR data [Boros et al., 2020]

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What works

- NER can handle (reasonably) noisy OCR data [Boros et al., 2020]
- NER can detect contaminated forms [Koudoro-Parfait and Lejeune, 2022]
  - . . . but still we will have OOV words to deal with

Roads to improvement for contaminated entities :

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What works

- NER can handle (reasonably) noisy OCR data [Boros et al., 2020]
- NER can detect contaminated forms [Koudoro-Parfait and Lejeune, 2022]
  - . . . but still we will have OOV words to deal with

Roads to improvement for contaminated entities :

- **Correction** post-OCR does not help much [Petkovic and Koudoro-Parfait]

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What works

- NER can handle (reasonably) noisy OCR data [Boros et al., 2020]
- NER can detect contaminated forms [Koudoro-Parfait and Lejeune, 2022]
  - . . . but still we will have OOV words to deal with

Roads to improvement for contaminated entities :

- **Correction** post-OCR does not help much [Petkovic and Koudoro-Parfait]
- **Entity Linking** with French data worsens the results

What we have seen

Beware of magical perfectly working examples (in particular in English [Bender, 2009, Bender, 2019])

What works

- NER can handle (reasonably) noisy OCR data [Boros et al., 2020]
- NER can detect contaminated forms [Koudoro-Parfait and Lejeune, 2022]
  - . . . but still we will have OOV words to deal with

Roads to improvement for contaminated entities :

- **Correction** post-OCR does not help much [Petkovic and Koudoro-Parfait]
- **Entity Linking** with French data worsens the results
- **Entity matching** (Alignment) is more promising [Koudoro-Parfait et al., 2022]

References I


Bender, E. M. (2009).
Linguistically naïve != language independent: Why NLP needs linguistic typology.
In *Proceedings of the EACL 2009*, ILCL '09, pages 26–32. ACL.


Bender, E. M. (2019).
#BenderRule: On naming the languages we study and why it matters. The Gradient.


Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N., and Doucet, A. (2020).
Robust named entity recognition and linking on historical multilingual documents.
In *CLEF 2020*, volume 2696, pages 1–17. CEUR-WS Working Notes.


Koudoro-Parfait, C. (2022).
Evaluating clustering for aligning contaminated forms of ne in noisy ocr data (in french).
Workshow Robustness of NLP systems, Caico Corro and Gaël Lejeune editors, page 5.


Koudoro-Parfait, C. and Lejeune, G. (2022).
Spatial NER on noisy literary corpus : form entities to maps (in French).
In *Séminaire des sources aux Systèmes d'Information Géographique*.


Koudoro-Parfait, C., Lejeune, G., and Buth, R. (2022).
Ner on noisy ocr data : proposals for automated morphological disambiguation (in French).
In *TAL-HN @ TALN(Traitement Automatique des Langues Naturelles) 2022*.