Discovering Spatial Relations in Literature: What is the Influence of OCR noise ?

Caroline Parfait (1,2,3) Gaël Lejeune (2)
Motasem Alrahabi (1,3) Glenn Roe (1,3)

The NewsEye International Conference

March 2021

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

Table of Contents

Introduction

Phd Subject

Literary space analysis, three dimensions

Data
- Literature VS News
- OCR data quality and its influence

Introduction

Phd Subject

Literary space analysis, three dimensions

Data
- Literature VS News
- OCR data quality and its influence

Task/users
- What do the users want ?
- Do we evaluate it properly with P/R/F-score ?

Introduction

Phd Subject

Literary space analysis, three dimensions

Data	• Literature VS News • OCR data quality and its influence
Task/users	• What do the users want ? • Do we evaluate it properly with P/R/F-score ?
Methods	• Adaptability of Named-Entity Recognition systems • Complementarity between these systems

Introduction

Phd Subject

Literary space analysis, three dimensions

Data	• Literature VS News • OCR data quality and its influence
Task/users	• What do the users want ? • Do we evaluate it properly with P/R/F-score ?
Methods	• Adaptability of Named-Entity Recognition systems • Complementarity between these systems

- How to overcome the limitations of NER quality ?
- Do NER tools meet the expectation of users ?

Introduction

Proposed approach : "user-centred design".

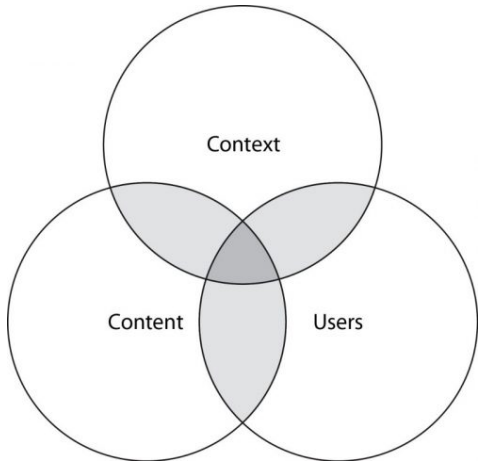
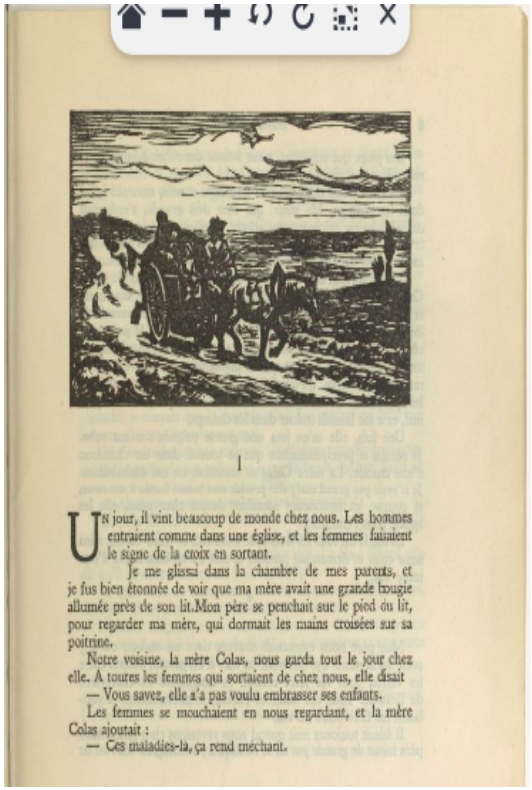


Figure: Peter Morville, *"Three Circles of Information Architecture"*, 2004

- **Task** : Spatial NER in **imperfect conditions**
- **Users** : Researchers in **various fields**
- **Data** : A rather heterogeneous corpus of French novels (19th/20th cent.) $\rightarrow$ **variability** ?

Correlation between input quality and output quality ?



I

Un jour, il vint beaucoup de monde chez nous. Les hommes entraient comme dans une église, et les femmes faisaient le signe de la croix en sortant.

Je me glissai dans la chambre de mes parents, et je fus bien étonnée de voir que ma mère avait une grande bougie allumée près de son lit. Mon père se penchait sur le pied du lit, pour regarder ma mère, qui dormait les mains croisées sur sa poitrine.

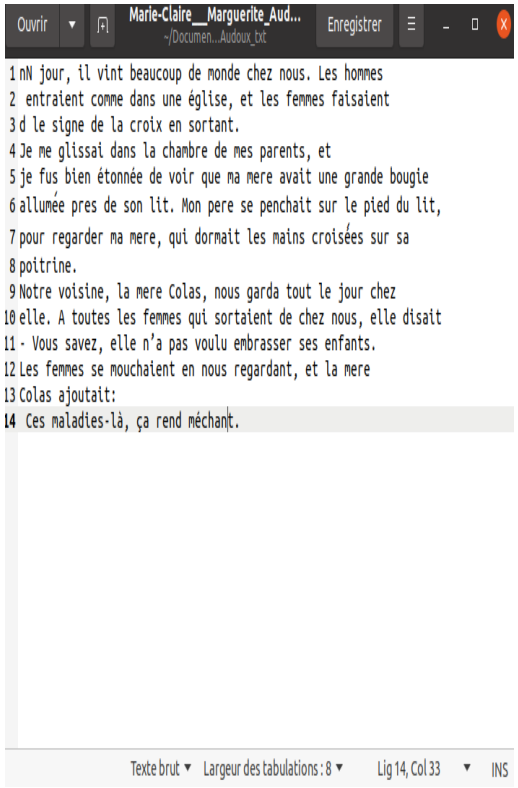
Notre voisine, la mère Colas, nous garda tout le jour chez elle. A toutes les femmes qui sortaient de chez nous, elle disait

— Vous savez, elle n'a pas voulu embrasser ses enfants.

Les femmes se mouchaient en nous regardant, et la mère Colas ajoutait :

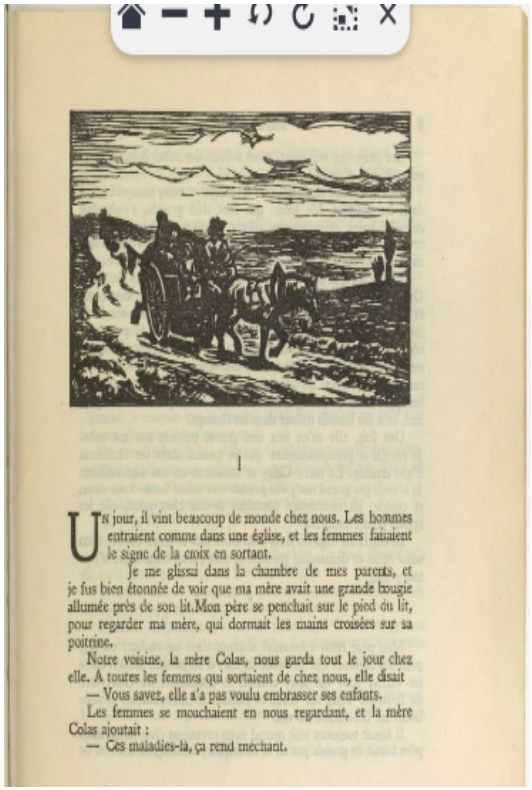
— Ces maladies-là, ça rend méchant.

(a) Image from Gallica



(b) Text after OCR processing

Correlation between input quality and output quality ?



I

Un jour, il vint beaucoup de monde chez nous. Les hommes entraient comme dans une église, et les femmes faisaient le signe de la croix en sortant.

Je me glissai dans la chambre de mes parents, et je fus bien étonnée de voir que ma mère avait une grande bougie allumée près de son lit. Mon père se penchait sur le pied du lit, pour regarder ma mère, qui dormait les mains croisées sur sa poitrine.

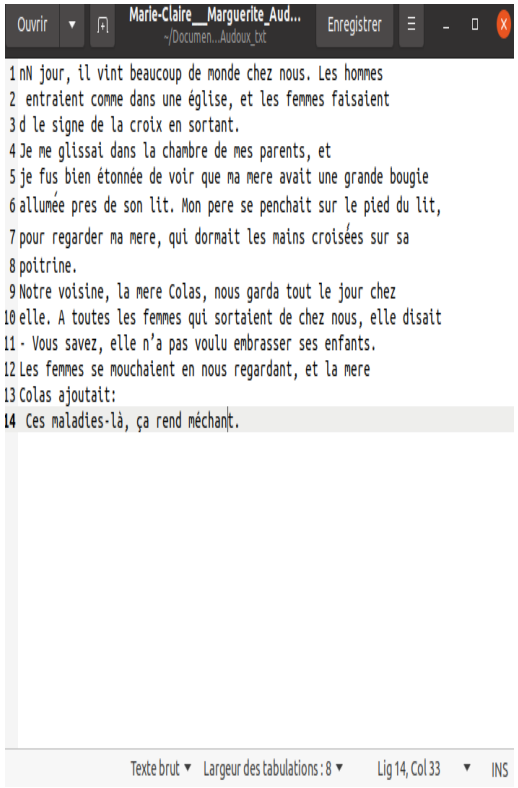
Notre voisine, la mère Colas, nous garda tout le jour chez elle. A toutes les femmes qui sortaient de chez nous, elle disait

— Vous savez, elle n'a pas voulu embrasser ses enfants.

Les femmes se mouchaient en nous regardant, et la mère Colas ajoutait :

— Ces maladies-là, ça rend méchant.

(a) Image from Gallica



(b) Text after OCR processing

- Distance in inputs (ELTeC version VS OCR)
- Distance in outputs (NER on ELTeC version VS NER on OCR data)?
- Causality between noise in input and noise in output ?

Explicit comparisons for humans

Data : "Je suis un aventurier", J.Dutronc (NER with Spacy SM)

- Exp. 1 : NER output for two texts with many NE
- Exp. 2 : Comparisons of distances between input and the NER output

Explicit comparisons for humans

Data : "Je suis un aventurier", J.Dutronc (NER with Spacy SM)

- Exp. 1 : NER output for two texts with many NE
- Exp. 2 : Comparisons of distances between input and the NER output

Clean Text	**NER**	**Noisy text**	**NER**
J'ai fait la vie à Varsovie	Varsovie	J'ai fait la vie Varsovie.	Varsovie (TP)
J'ai fait le rat à Canberra	Canberra	J'ai fait le rat a Camberra.	Camberra (FP?)
J'ai fait des games à Birmingham	Birmingham	J'ai fait des games à Binningham.	'**PER**' (FP)

Explicit comparisons for humans

"Les trappeurs de l'Arkansas (4e édition) ",Gustave Aimard, 1858 .

Clean Text	**NER**	**OCR Text**	**NER**
Le voyageur qui pour la première fois débarque dans l'Amérique du Sud éprouve malgré lui un sentiment de tristesse indéfinissable	(Amérique du Sud, 'LOC')	Le voyageur qui pour la premiere fois **débarquedans** l'Amerique du Sud, éprouve malgré lui unsentiment de tristesse indefinissable	(du Sud, 'LOC')
...	...	...	...

Explicit comparisons for humans

"Les trappeurs de l'Arkansas ",Gustave Aimard, 1858 .

Clean Text	**NER**	**OCR Text**	**NER**
C'est à ce déplacement continuel qu'il faut attribuer, en Amérique, l'absence de ces grands monuments,	Amérique	C'est à ce déplacement continuel qu'il faut attri buer, en Amérique, l'absence de ces grands monu ments,	Amérique
... et ces milliers d'animaux que la civilisation des autre parties de **l'Amérique** refoule de jour en jour...	**()**	... et ces milliersd'animaux que la civilisation des autre parties **del'Amerique** refoule de jour en jour...	**del'Amerique**
...	...	...	...

Explicit comparisons for humans

Some results

Can the user be pleased/confident with the output?

- Despite spelling mistakes the tool recognizes a spatial NE
- The system uses Syntactic information, not only the lexicon

Distances and similarities: Automatic comparisons

"Je suis un aventurier" Jacques Dutronc, 1970

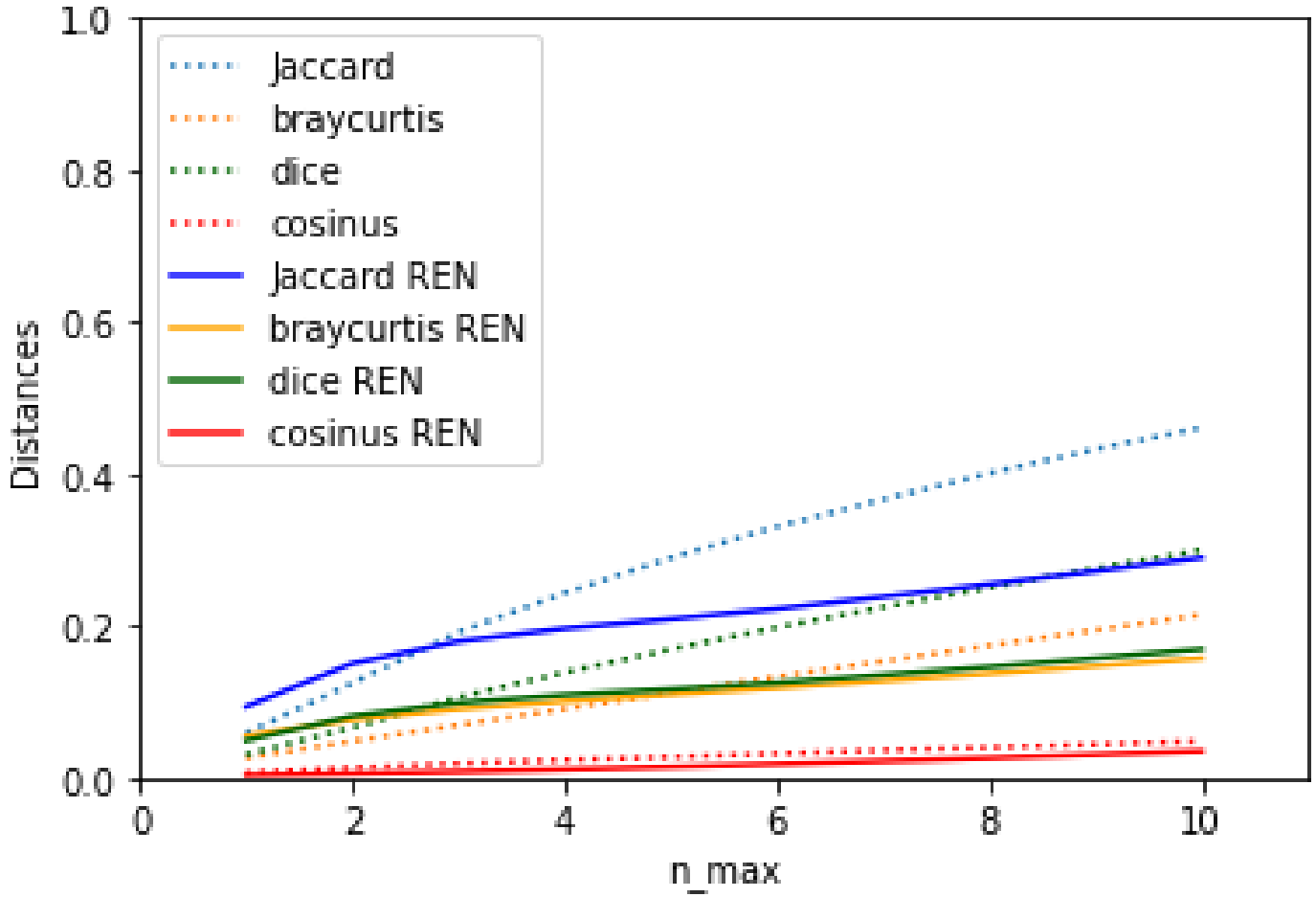


Figure: Distances in the input ("Official text" VS OCR **WER = 0.234**) and in the output (NER on clean data VS NER on OCR **WER=0.25**)

Distances and similarities: Automatic comparisons

Les trappeurs de l'Arkansas, Gustave Aimard, 1858.

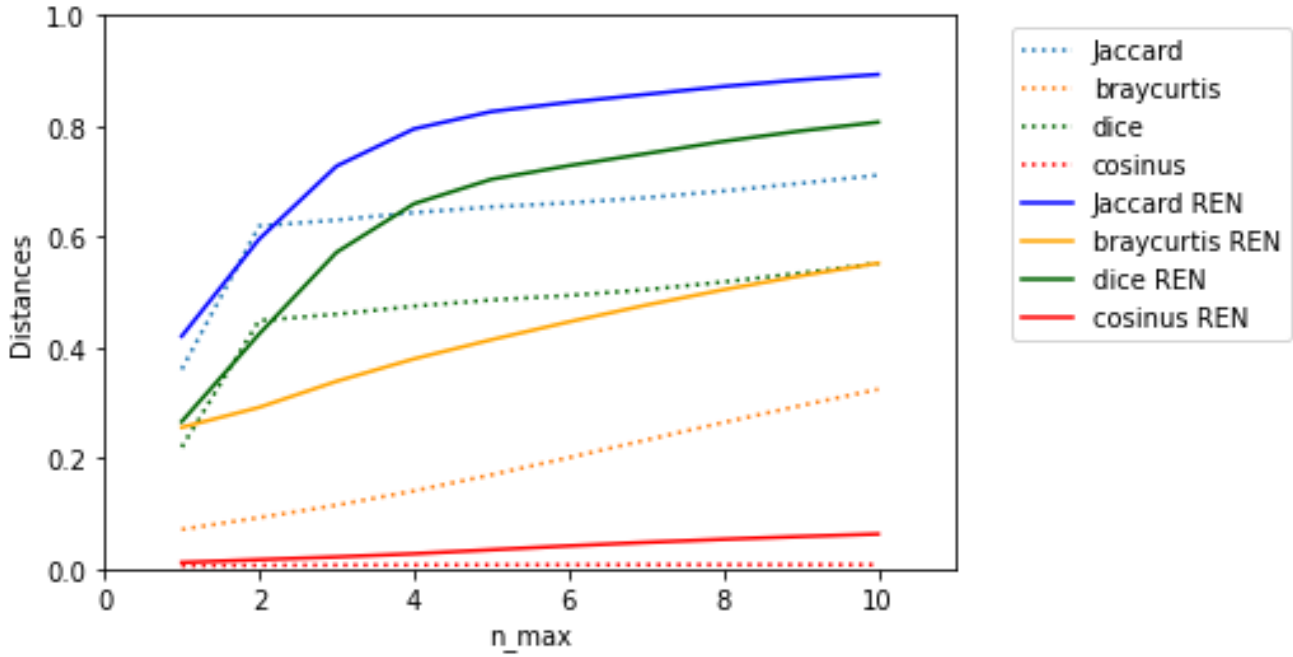


Figure: Distances betweens the inputs (Clean VS OCR **CER = 1.11**) and outputs (NER on clean data VS NER on OCR **CER = 0.9**)

Conclusion and Perspectives

What have learnt so far :

- OCR is the problem ?
- Recurring OCR errors that interfere with NER
- Glass Ceiling in NER [Stanislawek et al., 2019]

Conclusion and Perspectives

What have learnt so far :

- OCR is the problem ?
- Recurring OCR errors that interfere with NER
- Glass Ceiling in NER [Stanislawek et al., 2019]

Some solutions :

- Train models on noisy data ?
- Post-process the NER output rather than input from OCR (parsimonious?)

References I


Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., and Biecek, P. (2019).
Named entity recognition - is there a glass ceiling?
In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633.