

Pseudonymisation de documents textuels

Le cas des décisions de justice

Pavel SORIANO-MORALES

Etalab Data Drink #4

February 21, 2019

DINSIC – ETALAB

Qu'est-ce que la pseudonymisation ?

Qu'est-ce que la pseudonymisation ?

Exemple JURITEXT000025735516

RÉPUBLIQUE FRANÇAISE
AU NOM DU PEUPLE FRANÇAIS

COUR D'APPEL DE BASSE-TERRE

CHAMBRE SOCIALE

ARRET No 153 DU SEIZE AVRIL DEUX MILLE DOUZE

APPELANTE

Madame Rosine DU-NOM-FAUX

123, rue de la Vieille-Lanterne

97130 CAPESTERRE BELLE EAU

née le 01 janvier 2000

Représentée par Me Christiane ROMIL (TOQUE 119) [...]

Qu'est-ce que la pseudonymisation ?

Exemple JURITEXT000025735516

RÉPUBLIQUE FRANÇAISE
AU NOM DU PEUPLE FRANÇAIS

COUR D'APPEL DE BASSE-TERRE

CHAMBRE SOCIALE

ARRET No 153 DU SEIZE AVRIL DEUX MILLE DOUZE

APPELANTE

Madame Rosine DU-NOM-FAUX

123, rue de la Vieille-Lanterne

97130 CAPESTERRE BELLE EAU

née le 01 janvier 2000

Représentée par Me Christiane ROMIL (TOQUE 119) [...]

Qu'est-ce que la pseudonymisation ?

Exemple JURITEXT000025735516

RÉPUBLIQUE FRANÇAISE
AU NOM DU PEUPLE FRANÇAIS

COUR D'APPEL DE BASSE-TERRE

CHAMBRE SOCIALE

ARRET No 153 DU SEIZE AVRIL DEUX MILLE DOUZE

APPELANTE

Madame B... D...

...

...

Représentée par Me E... F... (TOQUE 119) [...]

Pourquoi ?

- Les décisions doivent être mises à disposition du public dans le respect de la vie privée (*Loi pour une République numérique*)

Pourquoi ?

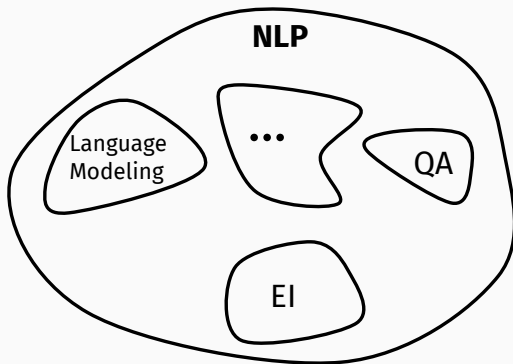
- Les décisions doivent être mises à disposition du public dans le respect de la vie privée (*Loi pour une République numérique*)
- Près de 3,9 millions de décisions de justice par an

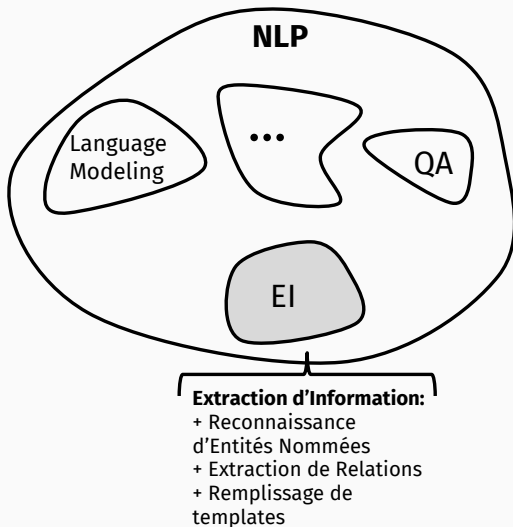
Pourquoi ?

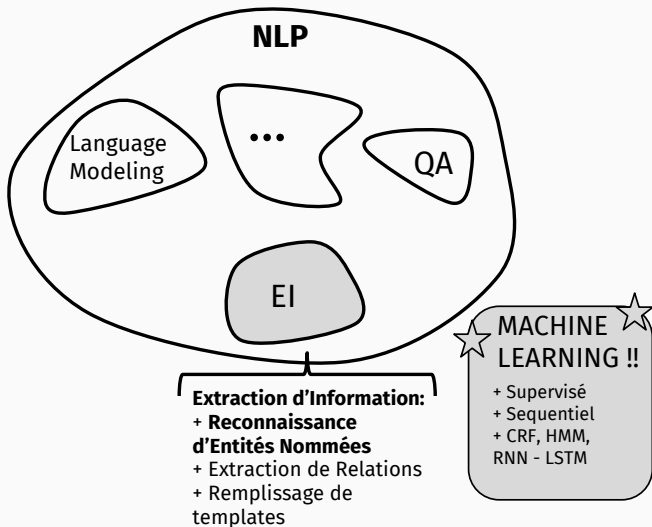
- Les décisions doivent être mises à disposition du public dans le respect de la vie privée (*Loi pour une République numérique*)
- Près de 3,9 millions de décisions de justice par an
- La relecture reste très coûteuse

Pourquoi ?

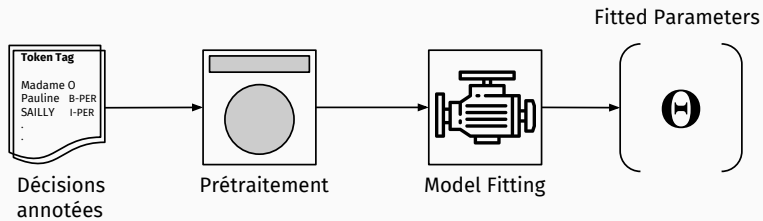
- Les décisions doivent être mises à disposition du public dans le respect de la vie privée (*Loi pour une République numérique*)
- Près de 3,9 millions de décisions de justice par an
- La relecture reste très coûteuse
- Besoin d'une solution automatisée



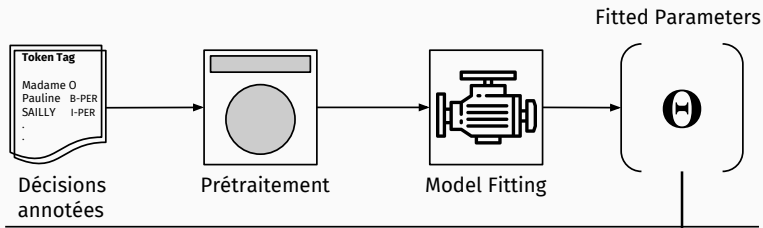




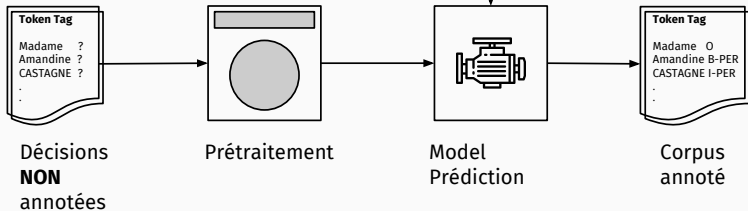
ENTRAÎNEMENT



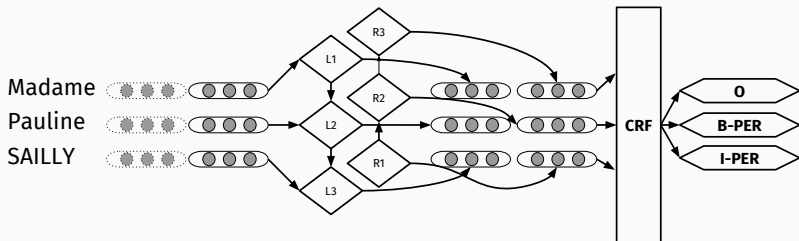
ENTRAÎNEMENT



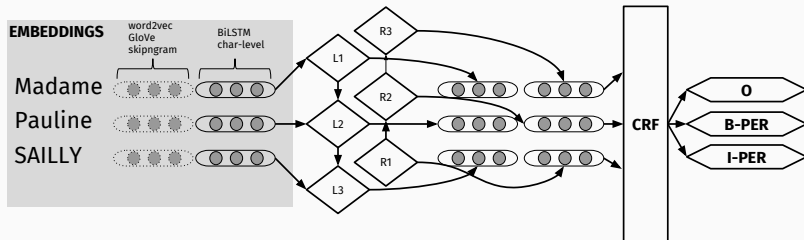
PRÉDICTION



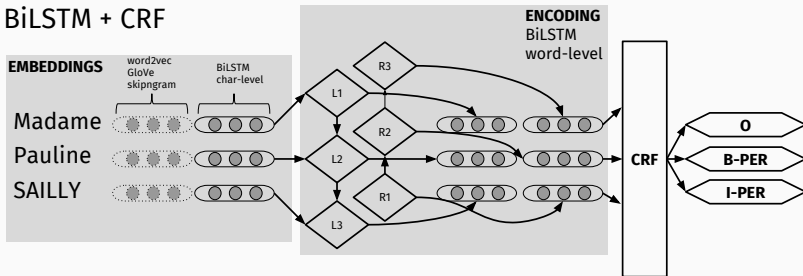
BiLSTM + CRF



BiLSTM + CRF

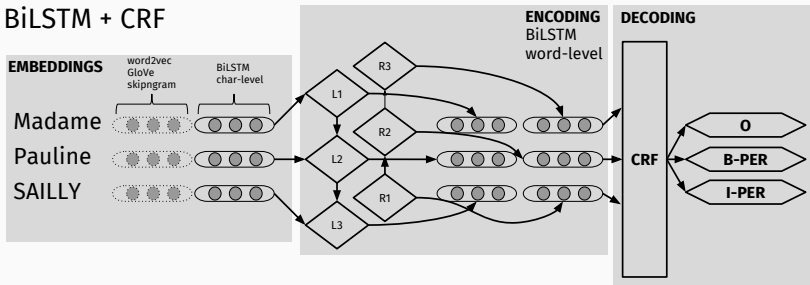


BiLSTM + CRF



L'algorithme

BiLSTM + CRF



Détails du corpus

Dataset	# de décisions	# de phrases	Tokens
train	57	6 989	173 448
dev	17	2 257	60 293
test	20	1 963	42 964
Totals	94	11 209	276 705

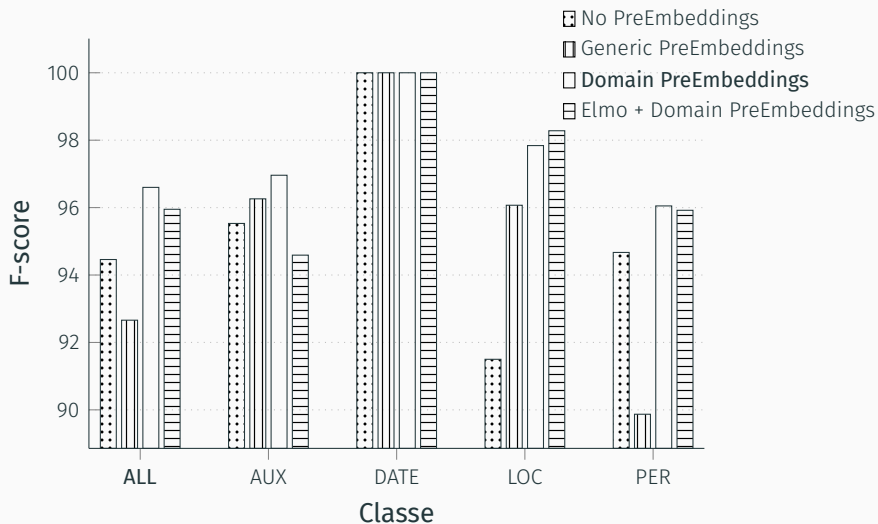
Dataset	AUX	DATE	LOC	PER	ALL
train	1 838	157	1 780	2 987	6 762
dev	562	54	566	848	2 033
test	545	26	434	629	1 634
Totals	2 935	237	2 780	4 464	10 429

AUX: avocats, membres de la formation de jugement

DATE: dates de naissance **LOC:** adresses de résidence

PER: parties et témoins.

Résultats des expériences



Pseudonymization Test

https://pseudo.etalab.studio

PSEUDONYMIZER!

Représenté par Me Florian DELISLE , avocat au barreau de LYON

Assisté de Me Pierre BESNARD, avocat au barreau de ROUEN

INTIMÉES

Pseudonymize!

PSEUDONYMIZED TEXT

Représenté par Me ... , avocat au barreau de LYON

Assisté de Me ... , avocat au barreau de ROUEN

INTIMÉES

SARL MOS AMBULANCE

ayant son siège social ...

...

Représentée par Me ... , avocat au barreau de DOUAI

Assistée de Me ... , avocat au barreau de REIMS, substitué par Me ... , associé

TAGGED TEXT

Représenté par Me Florian DELISLE, avocat au barreau de LYON

Assisté de Me Pierre BESNARD, avocat au barreau de ROUEN

INTIMÉES

SARL MOS AMBULANCE

ayant son siège social 99 rue Brice Arsenault BP 53333

69008 LYON

Représentée par Me Gérôme COURTET, avocat au barreau de DOUAI

Assistée de Me Léonard HENNEQUIN, avocat au barreau de REIMS, substitué par Me COULOMB, associé

What now ?

Améliorer le modèle/système

- Post-traitement à base de règles
- Feature Engineering
- Corriger les erreurs orthographiques
- Tester des autres algos

Obtenir plus de données

- Génération de données synthétiques (fake data)
- Profiter des décisions déjà pseudonymisées
- Possibilité d'entraîner et tester sur un corpus beaucoup plus large

Détection des pseudonymes (Regex/ML)

Le contrat de travail de Mme X... YYY née le [...] demeurant xxxxxxxx, passant à temps partiel sur une base de 20 heures

Classification des pseudonymes (ML)

Le contrat de travail de Mme **PER PER** née le **DATE** demeurant **LOC**, passant à temps partiel sur une base de 20 heures

Remplacement des pseudonymes (Aléatoire)

Le contrat de travail de Mme **Marie DUPONT** née le **01 janvier 2018**,
demeurant **99 rue Raoul Servant 69007 LYON**, passant à temps
partiel sur une base de 20 heures

Conclusion

- La tâche semble faisable avec des méthodes du Machine Learning / NLP
- Le nettoyage du texte et le post-traitement semblent très importants
- Plus des données sont nécessaires pour valider nos approches

Merci !

Des questions ?

github.com/psorianom

twitter.com/psorianom

Analyse d'erreurs : AUX et LOC

Classe AUX			Classe LOC		
Token	Real	Predicted	Token	Real	Predicted
signé	O	O	Unité	B-LOC	O
parMadame	O	O	Sud	I-LOC	O
CALOT*	I-AUX	O	Secteur	I-LOC	O
Conseiller	O	O	2	I-LOC	O
en	O	O			
l'	O	O	Centre	B-LOC	B-LOC
absence	O	O	Hospitalier	I-LOC	I-LOC
de	O	O			

Analyse d'erreurs : PER

Token	Classe PER	
	Real	Predicted
M.	O	O
Julien	B-PER	B-PER
Chavane	I-PER	I-PER
de	I-PER	O
Roissy	I-PER	O

