BookNLP-fr, the French Versant of BookNLP

A Tailored Pipeline for 19th and 20th Century French Literature

Frédérique Mélanie-Becquet* & Jean Barré* & Olga Seminck* & Marco Naguib & Martial Pastor & Clément Plancq & Thierry Poibeau*

1er octobre 2024

*Lattice Lab : ENS-PSL-CNRS

BookNLP-fr - A tailored pipeline for french fiction

Multilingual project at Berkeley

- French part of a multilingual project (Bamman, 2021)
- Trained on fr-Litbank, an annotated corpus of 19th and 20th century novels

Main tasks

- Literary NER (PER, FAC, TIME, ORG, LOC)
- Coreference Resolution with CamemBERT : pronouns, nouns, proper nouns : Arsène, **his** friend, **the gentleman thief**
- Clustering characters :
  Arsène Lupin, Monsieur Lupin, Lupin –> ARSENE_LUPIN

BookNLP-fr : Output example

Figure 1 : Coreference chains in Sarrasine (Balzac, 1830)

Annotated corpus

Year	Author	Title	
1830	Balzac	La maison du chat qui pelote	Full Text
1830	Balzac	Sarrasine	D10K
1836	Gautier	La morte amoureuse	D10K
1837	Balzac	La maison Nucingen	Full Text
1841	Sand	Pauline	D10K
1856	Cousin	Madame de Hautefort	D10K
1863	Gautier	Le capitaine Fracasse	D10K
1873	Zola	Le ventre de Paris	D10K
1881	Flaubert	Bouvard et Pécuchet	D10K
1882-1883	Maupassant	Mademoiselle Fifi (1)	D10K
1882-1883	Maupassant	Mademoiselle Fifi (2)	D10K
1882-1883	Maupassant	Mademoiselle Fifi (3)	D10K
1901	Achard	Rosalie de Constant	D10K
1903	Conan	Élisabeth Seton	D10K
1904-1912	Rolland	Jean-Christophe (1)	D10K
1904-1912	Rolland	Jean-Christophe (2)	D10K
1917	Bourgeois	Némoville	D10K
1923	Radiguet	Le diable au corps	D10K
1926	Audoux	De la ville au moulin	D10K
1937	Audoux	Douce Lumière	D10K

Table 1 : Short stories and novels in fr-Litbank

Entities	Number of occurrences
PER - Mentions	32,338
PER - Characters	3,006
FAC	2,325
TIME	1,836
LOC	1,040
GPE	928
VEH	475
ORG	205
TOTAL	**39,147**

Table 2 : Number of occurrences per type of entity.

Software development

BookNLP-* : BiLSTMs & CamemBERT

- A BiLSTM-CRF for entity recognition
- A BiLSTM-feed forward model for coreference resolution
- CamemBERT - a BERT based model tailored for French - Contextual embeddings
- Post-prod algorithms to get character sheet informations : actions, adjectives, age, gender, family, synsets, ..

	precision	recall	F_1
PER	85.0	92.1	88.4
LOC	59.4	54.3	56.8
FAC	73.4	66.0	69.5
TIME	75.3	36.4	49.1
VEH	68.9	63.6	66,1
GPE	68.2	52.9	59,6

Table 3 : NER evaluation, fr-Litbank

Metrics	F_1	
MUC	88,0	*Average* 76.4
B^3	69,2	
$CEAF_e$	71.8	

Table 4 : Coreference evaluation, fr-Litbank

Use case for BookNLP-fr

Classifying subgenres

- Subset of the Chapitres corpus - 650 novels (1850-1950)
- 5 subgenres - Children - Memoirs - Detective - Adventure - Romance

Research Questions

- Can BookNLP based features be a more interpretable substitute to BoW / Topics approach ?
- Is character information sufficient to classify subgenres ?
- Which feature is contributing more to detect adventure novels ?

Methods - Textual features + Subgenre classification

BoW features

- **Lexicon** : The N most frequent words.

VS

BookNLP-fr features

- **Characterization** : Adjectives and verbs that describe characters in the narrative. (ADJ + AGENT + PATIENT)
- **Chronotope** : LOC + FAC + VEH + TIME

Modeling

- Doc2vec representation for each facet
- **SVM** : Supervised model

Evaluation of models - BookNLP-fr

	F1-score BoW	F1-score BookNLP-fr	Support
Children	0.75	0.71	130
Memoirs	0.80	0.84	130
Detective	0.67	0.70	130
Adventure	0.62	0.73	130
Romance	0.80	0.75	130
Full Dataset	**0.72**	**0.75**	650

Table 5 : Classification Report BoW vs BookNLP-fr features

Confusion matrix - BoW & BookNLP-fr

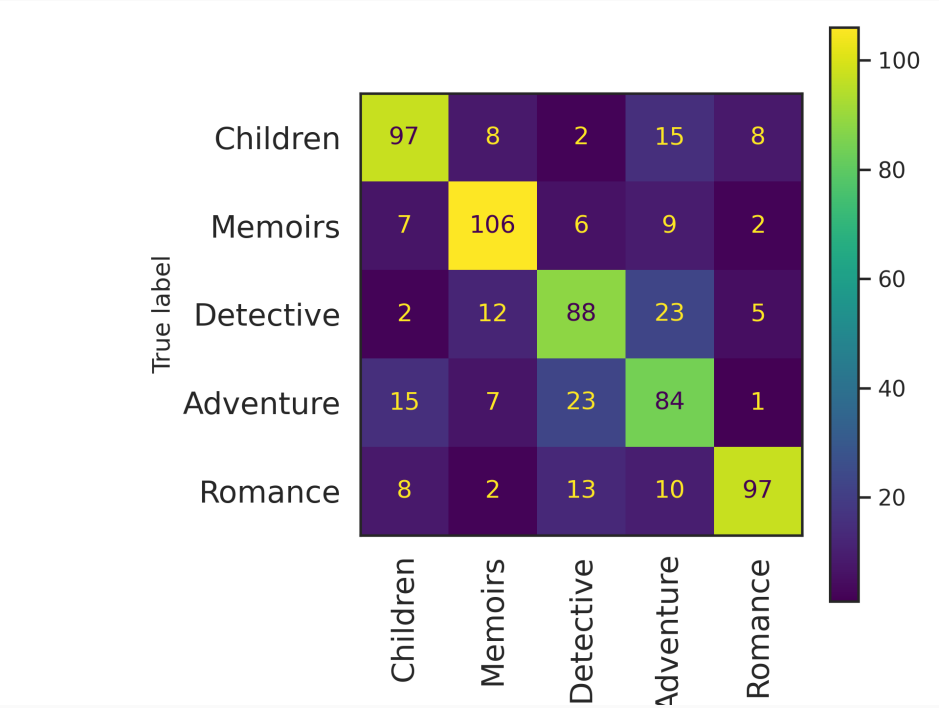

Figure 2 : Confusion Matrix for BoW features

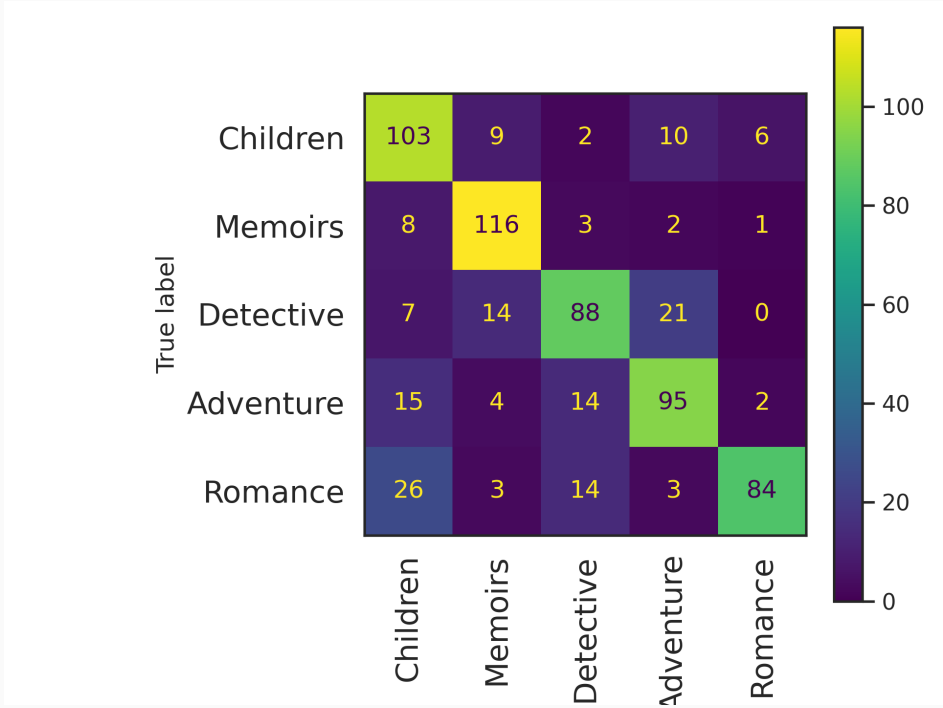

Figure 3 : Confusion Matrix for BookNLP-fr features

Evaluation of BookNLP-fr facets for classification

BookNLP-fr features	Accuracy
LOC	0.45
FAC	0.59
VEH	0.42
GPE	0.47
TIME	0.50
PATIENT	0.52
AGENT	0.62
ADJ	0.50
Baseline	0.2

Table 6 : Classification accuracy for each BookNLP facet

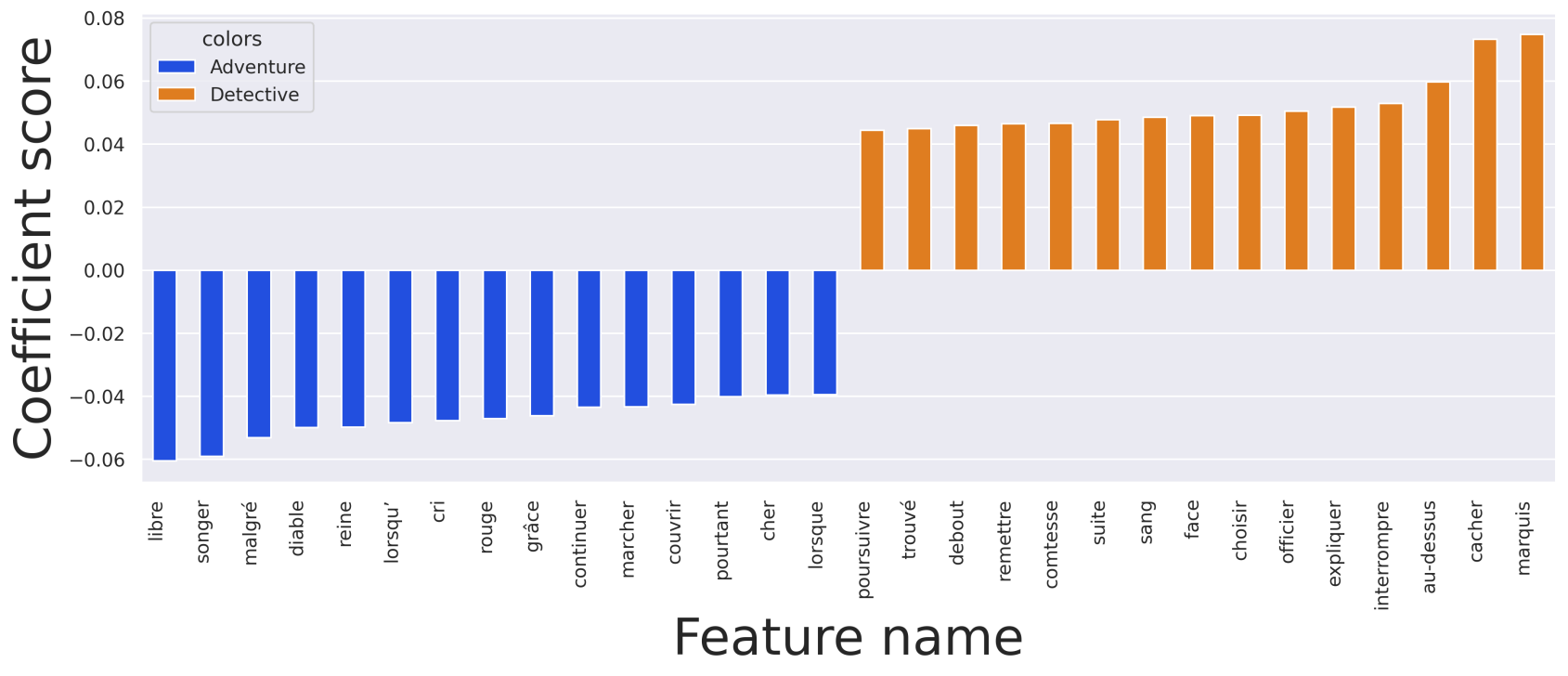

Figure 4 : BoW discriminant features for Adventure vs Detective classification.

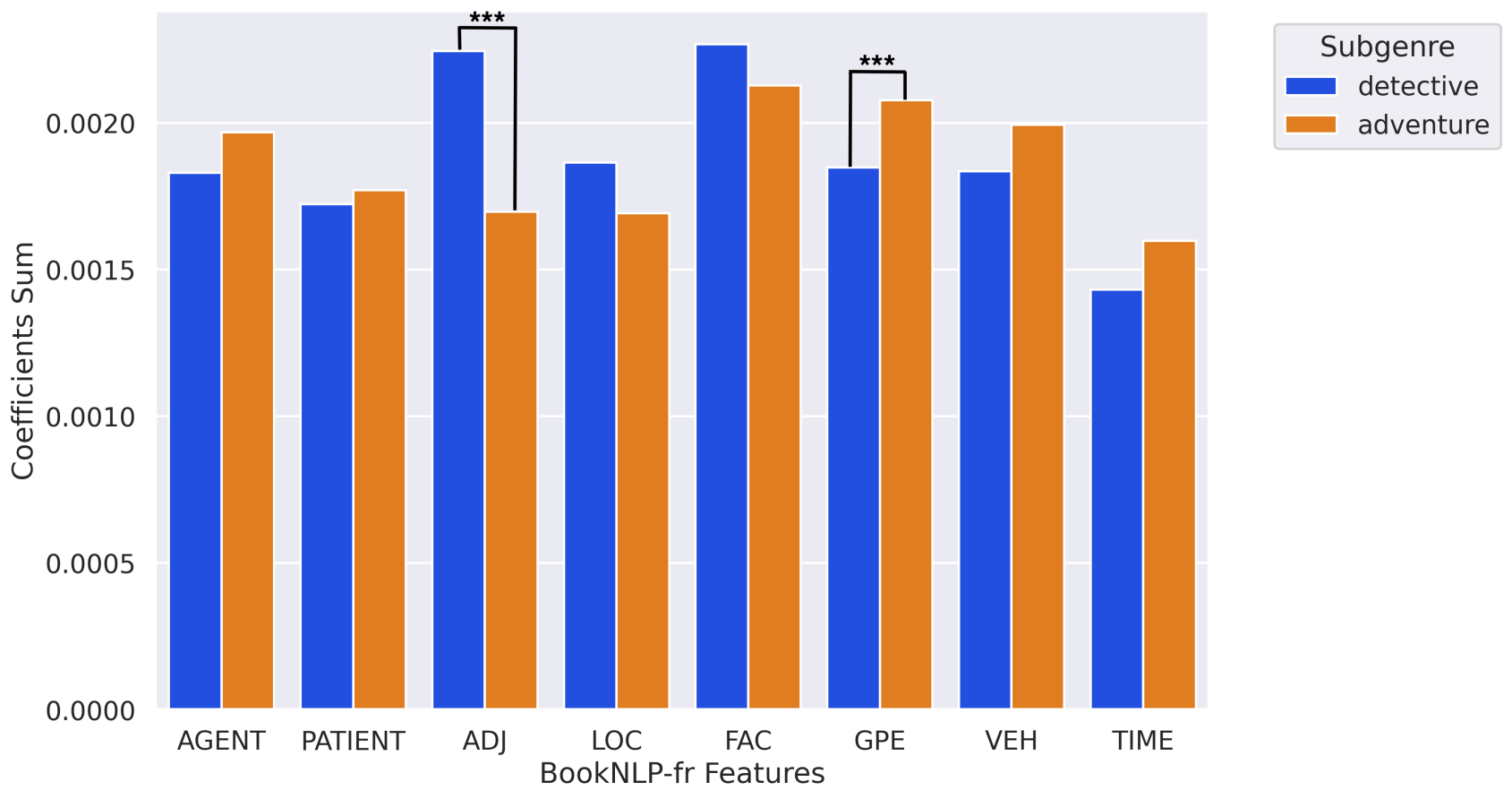

Figure 5 : BookNLP-fr discriminant features for Adventures vs Detective classification. *** meaning p<0.001.

Perspectives

BookNLP-fr - a distant reading ready pipeline

- Semantic view of novels
- Classifying subgenres
- Enables large scale study of characters
  Character networks - Character impact on Narratives - Evolution of characterization over time and subgenres

Limitations

- Still having some mistakes - Character splits or merges, lower recall
- A character specific evaluation metric ?

Thank you !

Questions ?

Feel free to reach us !

jean.barre@ens.psl.eu

thierry.poibeau@ens.psl.eu

Data & Code :

https://github.com/lattice-8094/fr-litbank

https://github.com/crazyjeannot/jcls_booknlp_subgenres

	BookNLP-fr			Camembert-NER		
pos_tag	precision	recall	F1 Score	precision	recall	F1 Score
PROP	82.5	79.2	**80.8**	91.85	72.05	80.75
NOM	74.9	74.7	**74.8**	96.32	14.17	24.70
PRON	86.3	89.5	**87.9**	100.00	0.10	0.20
ALL	82.39	83.88	**83.13**	92.58	7.92	14.59

Table 7 : PER recognition performance, BookNLP-fr vs Camembert-NER

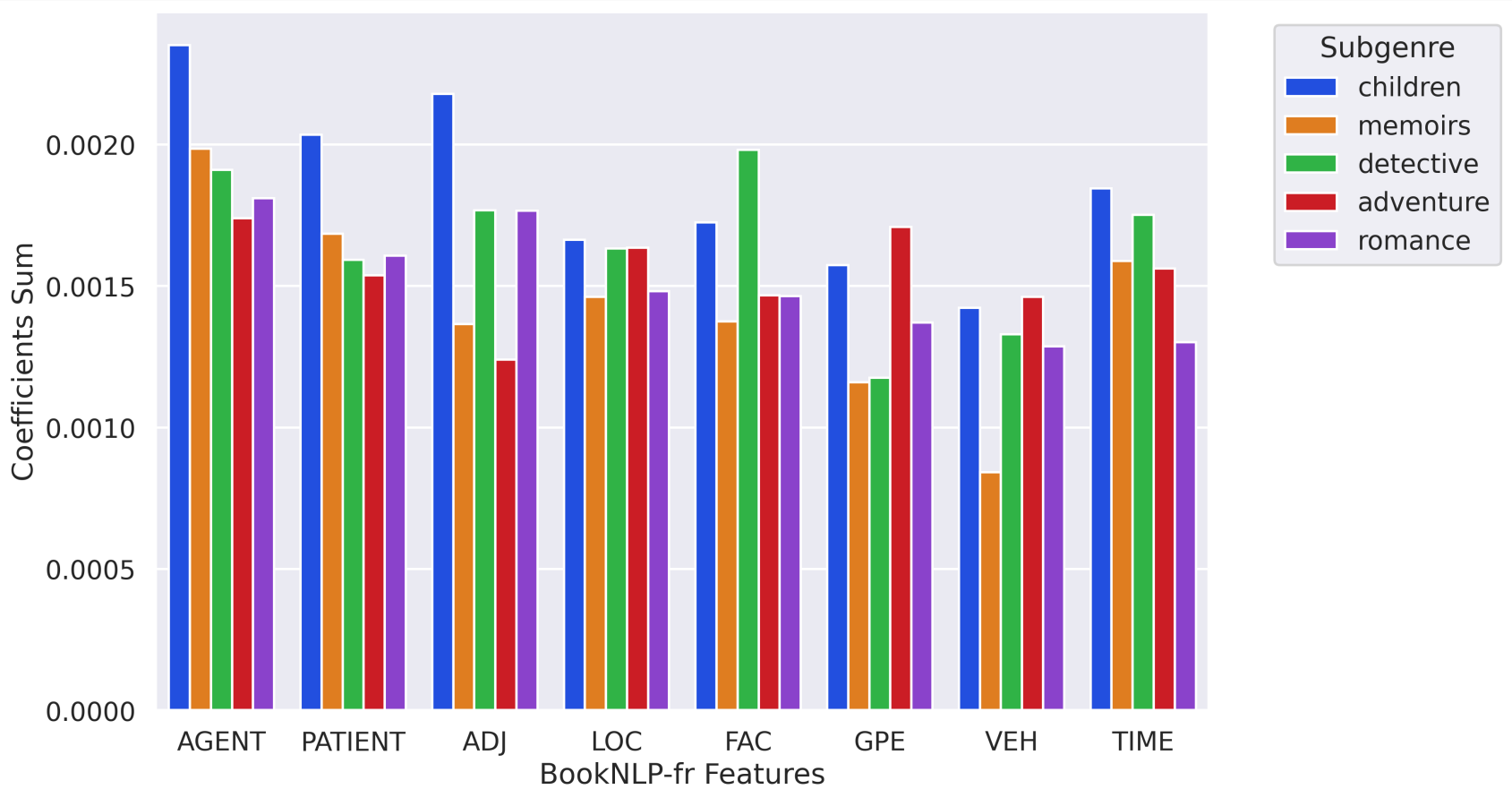

Figure 6 : BookNLP-fr discriminant features