Adapting a system for Named Entity Recognition and Linking for 19th century French Novels

Aicha Soudani, Yosra Meherzi, Asma Bouhafs, Francesca Frontini, Carmen Brando, Yoann Dupont, Frédérique Mélanie-Becquet

Adapting a system for Named Entity Recognition and Linking for 19th century French Novels

Aicha Soudani; Yosra Meherzi; Asma Bouhafs - ECSTRA, IHEC, Univ. de Carthage
Francesca Frontini - Praxiling UMR 5267, Univ. Paul-Valéry Montpellier 3
Carmen Brando - CRH UMR 8558 / EHESS
Yoann Dupont; Frédérique Mélanie-Becquet - Lattice (UMR8094), ENS, PSL University, Paris3 Sorbonne nouvelle

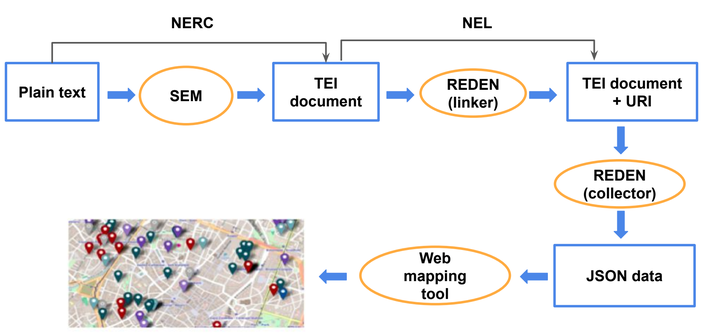
- **Annotation and linking of Named Entities in novels** - People, Places, or other proper names; important task for Digital Editions, Digital Literary Stylistics and Spatial Humanities
- **Create a pipeline for French Literary Texts of the 19th century**
  For "Paris Time Machine" Huma-Num consortium

Corpus:
- ❏ first two chapters of ***Le Ventre de Paris*** (Zola, 1873)
- ❏ first chapter of ***Cesar Birotteau*** (Balzac, 1837)
- ❏ the size is 30,889 words

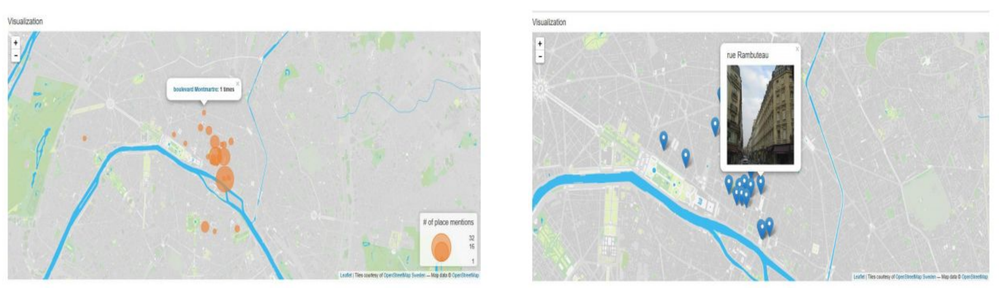
Named Entity Recognition and Classification with SEM (Dupont, 2017), a machine learning system based on conditional random fields (CRF)

Named Entity Linking with REDEN (Brando et al., 2016; Frontini et al., 2016):
- ❏ graph-based algorithm and Semantic Web technologies
- ❏ composed of two phases, candidate retrieval and disambiguation

The overall Pipeline



Visualisation output



By dereferencing IRIs, it is possible to access values for the aforementioned properties and use them in the context of a Web mapping application

SEM Evaluation

SETUP 1 - training on the first chapter of Zola and test on the second one
SETUP 2 - training on the Zola subcorpus and test on Balzac

	Setup 1	Setup 2
Precision	1	0,7
Recall	0,69	0,26
F-measure	0,82	0,38

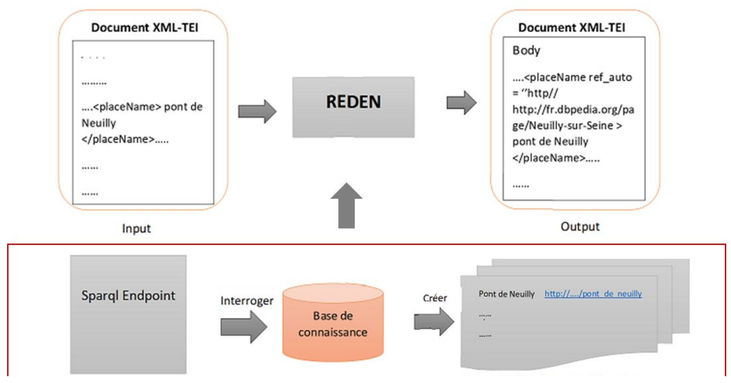
- ❏ The model trained on one chapter performs well on another; however we notice an important drop in performance when the model trained on one novel is applied on the other
- ❏ This bears important consequences for our ongoing work to constitute an adapted NERC model for 19th century French novels

The test and training set

Mentions	LVDP	CB
Persons	257	533
Places	187	41

- ❏ Inter-annotator agreement of 0,91 (F-measure)
- ❏ An Internationalized Resource Identifier (IRI) was added for those place names for which a reference existed.

REDEN Setup and Evaluation



KB	Overall linking accuracy
DBpedia	0,834
BNF	0,7
Wikidata	0,85

- ❏ REDEN was tested on the task of referencing *placeNames*, using three different KBs. Wikidata is the best KB in terms of overall accuracy
- ❏ BnF is a more accurate source of information for **old place names** (e.g. it records the older name "château de Bicêtre" for the "fort de Bicêtre")
- ❏ candidate retrieval may fail due to spelling variations

Results and future work

Annotated dataset is publicly available

We plan to increment the French corpus in connection to other initiatives
Labex OBVIL & COST ACTION Distant Reading for European Literary History - the French European Literary Text Collection (ELTeC)