French BookNLP Project

Progress report

Jean Barré & Thierry Poibeau

01 september 2022

Lattice Laboratory : ENS-PSL-CNRS

Table of Content

Annotation

Annotation

Build on Democrat Coref Corpus

- 18 french novels (10,000 first tokens) from the XIX^{e}
- 184,000 tokens - 14,208 ENTITIES annotated
- Co-ref annotation - Native from Democrat project

Statistics FR-LitBank

	train	*dev*	*test*	total
# tokens	15 9591	22 304	22 102	184 107
# chunks	9 974	1 394	1 381	11 506
# mentions PER	20 712	2 111	2 950	25 773
# entités PER	5 078	720	771	6 569

Table 1 : Statistics FR-LitBank

Event

Event annotation

- Integration of modals and negation in the annotation scheme
- ”He could not hold back his tears” : the character cried.
- Events : 14,305 events in the total dataset.

Training Pipeline

Training Pipeline

1. Fine-tune Camembert[1] (French Language Model)
2. BIOES (Beginning, Inside, Outside, Ending, Single-word) labeling scheme.
3. Overlapping chunks for long texts, (A new chunk starts every 16 tokens). Each words are in 32 chunks.

Results

Results Entities

precision	rappel	F_1
86.01	83,13	85,42

Table 2 : Test result FR-LitBank

Results Co-reference

		precision	rappel	F_1
Mentions		90,65	90,08	90,37
Coreference	*MUC*	85,06	85,10	85,08
	B^3	82,66	56,49	67,11
	$CEAF_e$	28,50	91,89	43,50
	BLANC	85,81	62,99	69,22
	LEA	64,73	62,47	63,58

Table 3 : Test result FR-LitBank

Results Events

precision	rappel	F_1
51.32	70,73	61,02

Table 4 : Test result FR-LitBank

Results Quotes

precision	rappel	F_1
91.95	90,74	91,34

Table 5 : Test result FR-LitBank

Issues & Futur work

Futur work

Already achieved :

- Provide 4 annotation models. Entities, Events, Coref, Quotes
- Provide a script to create your own model from your own annotated data

Futur work

- Evaluate how good / how bad our models are
- Long string (novels) behavior for Co-reference
- Quite an issue to build Character Networks

Bibliographie indicative i

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot.
CamemBERT : a tasty French language model.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

Data and code available on github :
https://github.com/lattice-8094/fr-litbank/