A Data-driven Approach to Named Entity Recognition for Early Modern French

Pedro Ortiz Suarez, Simon Gabay

A Data-driven Approach to Named Entity Recognition for Early Modern French

Pedro Ortiz Suarez[1], Simon Gabay[2]

[1]Data and Web Science Group - University of Mannheim,
[2]Digital Humanities Group - University of Geneva

Token	Lemma	POS	COARSE	FINE	FINE-COMP	NESTED	Wikidata ID
Les	le	Da	O	O	O	O	_
allemands	allemand	Nc	O	O	O	O	_
élurent	élire	Vvc	O	O	O	O	_
pour	pour	S	O	O	O	O	_
empereur	empereur	Nc	B-pers	B-pers.ind	B-comp.title	O	Q438435
Rodolphe	Rodolphe	Np	I-pers	I-pers.ind	B-comp.name	O	Q438435
duc	duc	Nc	I-pers	I-pers.ind	B-comp.title	O	Q438435
de	de	S	I-pers	I-pers.ind	I-comp.title	O	Q438435
Suabe	Souabe	Np	I-pers	I-pers.ind	I-comp.title	B-loc.adm.reg	Q438435

Tab. 1 - NERC Fine-Grained annotation with EL.

Abstract

Named entity recognition has become an increasingly useful tool for digital humanities research, specially when it comes to historical texts. However, historical texts pose a wide range of challenges to both named entity recognition and natural language processing in general that are still difficult to address even with modern neural methods. In this article we focus in named entity recognition for historical French, and in particular for Early Modern French (16th-18th c.), i.e. *Ancien Régime* French. However, instead of developing a specialised architecture to tackle the particularities of this state of language, we opt for a data-driven approach by developing a new corpus with fine-grained entity annotation, covering three centuries of literature corresponding to the early modern period; we try to annotate as much data as possible producing a corpus that is many times bigger than the most popular NER evaluation corpora for both Contemporary English and French. We then fine-tune existing state-of-the-art architectures for Early Modern and Contemporary French, obtaining results that are on par with those of the current state-of-the-art NER systems for Contemporary English. Both the corpus and the fine-tuned models are released.

Corpus

Rather than designing a new corpus, we have decided to use a subpart of the "core corpus" of the *Presto* project [1], namely the text written during the French *Ancien Régime* (c.15th-18th c., i.e. 34 texts). This choice is driven by our will to limit the number of annotated corpora for historical French, the same set of documents having already been abundantly corrected to train a lemmatizer, but also to avoid a complex selection of works supposed to ensure a relative representativeness of literary documents from the *Ancien Régime*, already perfectly done by our colleagues.

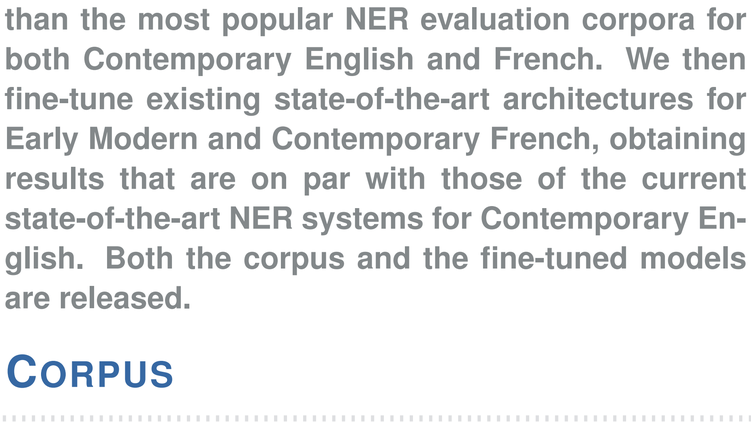

Fig. 1 - Number of tokens per century.

The number of genres covered is extremely large: poetry, drama, novel, correspondence, grammar, philosophy, short stories, encyclopedic literature, etc. and guarantees, here again, a reasonable representativeness of the range of possibilities of *Belles-Lettres*[1]. The corpus is balanced regarding the distribution per century (c. 10/century) but not regarding the length of the texts, which increases over time (cf. fig. 1), following a possible trend in literature.

Annotation

It seemed logical to follow the *Quaero* annotation guide, that is used by two important historical corpora presented *supra* (*Quaero* and *Impresso*). Because our texts and interests diverge from those of the aforementioned corpora, only some types and subtypes have been kept from the *Quaero* typology. The details of our choices can be found in a dedicated annotation manual.

The annotated texts are available in multi-columns `tsv` files (cf. tab. 1). Each token has a lemma (manually corrected) and a POS (produced by the *Presto* project, non-systematically corrected but fairly reliable) using the MULTEXT tag set. We propose a coarse-grained annotation for high-level entity types and fine-grained annotation using subtypes using the following syntax:

```
BIO-TYPE.SUBTYPE
For instance: B-loc.adm.town
```

Results

Table 2 shows a brief overview of our results, we can see that our BiLSTM-CRF already produces quite strong results, attaining an f1-score of 0.8586 which is quite remarkable taking into account how heterogeneous our corpus is and how different the data itself is from the pre-training data used in the FastText word embeddings of the Bi-LSTM model.

Model	Precision	Recall	F1-Score
BiLSTM-CRF	0.8640	0.8533	0.8586
CamemBERT	*0.9303*	*0.9309*	*0.9306*
D'AlemBERT	**0.9329**	**0.9323**	**0.9326**

Tab. 2 - Comparison between D'AlemBERT, CamemBERT and an LSTM-CRF-based model performance on the test set of our corpus, results are averaged over 10 runs with different seeds.

On the other hand for both CamemBERT and D'AlemBERT we obtain quite high results above the 0.93 in f1-score. These results are quite remarkable because in spite of how heterogeneous our corpus is, and despite of the challenges posed by an historical language previously discussed, we obtain results that are almost on par with the current state of the art architectures for Contemporary English.

For the CamemBERT and D'AlemBERT results by entity type (cf. tab. 3), we see results which actually supports our hypothesis that due to the size of our corpus, the Transformer-based models might be "*forgetting*" some of their pre-training contemporary data and "*re-learning*" the training data of our corpus seen during fine-tuning. There is a small exception to this and it again the *production* entity type, we can see that D'AlemBERT performs a bit better for this particular type which might be explained by the presence of these in D'AlemBERT's pre-training data as opposed to the lack of it in CamemBERT's web-based pre-training corpus, suggesting that while these models might be "*forgetting*" while exposed to corpora of the size of our corpus, they can still leverage their pre-training data to a certain extent.

CamemBERT					D'AlemBERT				
Entity Type	Precision	Recall	F1-Score	Support	Entity Type	Precision	Recall	F1-Score	Support
pers	0.9373	0.9236	0.9304	2734	pers	0.9355	0.9279	0.9317	2734
loc	0.9140	0.9371	0.9254	1384	loc	0.9242	0.9335	0.9288	1384
amount	0.9840	0.9840	0.9840	250	amount	0.9800	0.9800	0.9800	250
time	0.9447	0.9407	0.9427	236	time	0.9456	0.9576	0.9516	236
func	0.9209	0.9143	0.9176	140	func	0.9333	0.9000	0.9164	140
org	0.8364	0.9388	0.8846	49	org	0.8148	0.8980	0.8544	49
prod	0.7742	0.8889	0.8276	27	prod	0.8621	0.9259	0.8929	27
event	0.8333	0.8333	0.8333	12	event	0.8333	0.8333	0.8333	12
micro avg	0.9303	0.9309	0.9306	4832	micro avg	0.9329	0.9323	0.9326	4832
macro avg	0.8931	0.9201	0.9057	4832	macro avg	0.9036	0.9195	0.9111	4832
weighted avg	0.9307	0.9309	0.9307	4832	weighted avg	0.9331	0.9323	0.9327	4832
samples avg	0.8856	0.8856	0.8856	4832	samples avg	0.8893	0.8893	0.8893	4832

Tab. 3 - Results of CamemBERT and D'AlemBERT on the test set of our corpus by entity type. Results are averaged over 10 runs with different seeds.

Contact: pedro.ortiz@uni-mannheim.de