”A Novel of Character”: Towards the Automatic Annotation of Characters in a Large Corpus of French Novels

B Rabu, F Mélanie, Thierry Poibeau

B. Rabu, F. Mélanie and T. Poibeau

"A NOVEL OF CHARACTER": TOWARDS THE AUTOMATIC ANNOTATION OF CHARACTERS IN A LARGE CORPUS OF FRENCH NOVELS

Abstract. In this paper, we apply named entity recognition techniques to a corpus of literary texts, i.e. French novels from the 18th, 19th and 20th century. We obtain results that are usable but could be improved by using advanced annotation techniques. We discuss the use of active learning in this context, as well as the different applications that could be derived from this kind of annotation. In particular, we show that the automatic annotation of large literary corpora makes it possible to check whether traditional classifications exhibit specific structural patterns that could be identified automatically.

Keywords. Named Entity Recognition; Digital Humanities; Literature Analysis; Text Mining; Distant Reading

1. Introduction

The recent availability of large literary corpora in different languages has open new pathways for the study of literature. This approach is often called "distant reading" (Moretti, 2013) since corpora are then too large to be read directly and can only be accessed through specific tools that create a "distance" between the text and the reader. This approach has given birth to new research avenues and researchers are now able to observe tendencies over a large number of texts, instead of focusing on isolated observations concerning a few novels.

A specific research programme includes for example the investigation of the structure of novels, through the notion of "character": How central are the different characters of a novel? How do they interact with each other in the course of the novel? In other words, are there specific patterns that emerge from different novel traditions, from different period of times or from different subgenres? (Piper *et al.*, 2017)

There are now several tools available for different languages that are able to recognize person names in texts and, more generally, named entities like locations, artefacts or organizations. Named entity recognition is a well-established task (Poibeau, 2003), but existing tools are far from perfect: they make errors and need to be re-trained to reach acceptable performance on different corpora (Finkel et al., 2005). Their performance over literary texts also need to be properly evaluated, as they are generally trained on news or other kinds of Web sources (Poibeau & Kosseim, 2001).

In this paper, we propose an experiment on a corpus of French novels. We annotate person names, as well as other related text sequences (like titles, functions, or occupations) that can be used to refer to a character. The question of "what to annotate" is a highly complicated one, and we will just give a brief overview of our annotation principles below. We first present the corpus, then our annotation scheme and the tool we used for our experiments. We then present our results on the different novels, we discuss these results and conclude with some observations for future work.

2. The Corpus

For our study, we chose different novels from the 18th, 19th and 20th century. The choice is of course quite subjective as a large number of novels is directly available online in an electronic format. We wanted to get a balanced corpus among the three centuries considered.

Title	**Author**	**Publication date**	**Size (approx. # of words)**
De l'esprit des lois	Montesquieu	1748	65.000
Candide	Voltaire	1759	32.000
L'an 2440	L-S. Mercier	1771	93.000
Les liaisons dangereuses	P. C. de Laclos	1782	140.000
Les Rêveries du promeneur solitaire	J-J. Rousseau	1782	40.000
Notre-Dame de Paris	V. Hugo	1831	156.000
La Maison Nucingen	H. de Balzac	1838	34.000
Madame Bovary	G. Flaubert	1857	116.000
Alice au pays des merveilles	L. Carroll	French: 1869 Original: 1865	30.000
À l'ombre des jeunes filles en fleurs	M. Proust	1919	205.000

Les Faux-Monnayeurs	A. Gide	1925	115.000
La Gloire de mon père	M. Pagnol	1957	47.000

3. Annotation Principles

One of the most difficult part of the task is to define the entities that should be annotated. Some examples are easy to recognize and annotate, but lots of others are difficult.

Person names: Both fictive and real names can be found in novels. Person's names correspond to proper names like first names (*Odette*), last names (*Swann*) or a combination of both (*Odette Swann*). These proper names can be preceded with a title (*Madame de Crécy*, *M. de Norpois*), which can lead to complex noun phrases, especially with nobility titles (*Son Éminence monseigneur le cardinal de Bourbon*, in *Notre-Dame de Paris* from Victor Hugo). The same phenomenon is observed with function or occupation names (*le marquis de Norpois*, *le professeur Cottard*, or *l'abbé Frayssinous*). Texts often contain references to characters through their function, occupation or title, without mentioning any proper name, especially when the character has already been introduced or when there is no ambiguity left with just the title or the function mentioned (*le Principal*, *le Vicomte*). Generic groups of people are not annotated as they cannot be directly considered as characters (*les Anglais*, *les Parisiens*), but specific groups must be annotated (like *les Swann*, in Proust's *A l'ombre des jeunes filles en fleurs*). Other difficult cases are words like *God* or *the Divinity*, whose status is unclear.

Other entities: The software we used for the annotation by default also annotates other kinds of entities (location names, companies, *etc*.). This is of course interesting for the study of literary texts, especially location names since one could imagine a joint study of people (characters) and places. However, this is outside the scope of the present study and, in what follows, we will just focus on person names.

We cannot give all the details used for the annotation here, but the interested reader can refer to existing guidelines, for example the one proposed by Rosset *et al.* (2011) that offers valuable principles for French, especially to practically solve difficult cases. Other guidelines exist for other languages but the most important principle is to be consistent throughout the annotation phase, since a part of the decisions to take is subjective, as there is no formal distinction between named entities and other referential expressions in natural languages.

4. The Annotation Tool

We used a tool called SEM for our experiments (Dupont, 2018). SEM is an open piece of software, freely available online, and based on machine learning techniques. More specifically SEM is based on Wapiti (Lavergne *et al.*, 2010), a CRF toolbox (Conditional random Fields, Lafferty *et al.,* 2001). CRF are simpler than neural networks, and they obtain competitive results for the annotation of sequences. They are thus especially indicated for tasks like named entity recognition, since our goal is to recognize local and continuous sequences of texts (sequences without gaps). SEM can also very easily be trained using an annotation interface. Practically the end user can just annotate a few examples before training a new model that can be tested on new data, which is what we needed to do since our results will highly depend on the training phase using a representative sample of our corpus.

We have trained a new model from scratch for each century, but this is of course far from optimal since it would normally require annotating huge quantities of data to achieve reasonable performance. There is moreover a serious risk of overfitting since we train a new model for each novel / century. One solution to this problem would be to dynamically update an existing model based on new data. Recent machine learning techniques makes this approach possible, but it has not been explored yet in our context. The other approach consists in using active learning techniques to accelerate and optimize the annotation phase. In our case, unlabelled data is abundant but manually labelling is expensive. Learning algorithms can actively query the user for labels, making it possible to dynamically and automatically identify interesting examples for training, i.e. discriminative and ambiguous examples that the system cannot annotate directly (typically, because contradictory indices can be found in the context). This approach is for example the one already used by Prodigy, the annotation tool developed in relation with Spacy by Montani and Honnibal (2017).

5. Results and Discussion

The results are given in table 1. All the results are expressed using F-measure, the harmonic mean of precision (the percentage of sequences that are accurately recognized among what has been recognized by the system) and recall (the percentage of sequences actually recognized among all those that should have been recognized).

Annotation model	**18th century corpus**	**19th century corpus**	**20th century corpus**
18th century	**0,68**	0,63	0,68
19th century	0,61	**0,70**	0,67
20th century	0,62	0,69	**0,73**
All	**0.72**	**0,77**	**0,86**

Table 1. annotation results (all the results are expressed using F-measure).

We can make two main observations: *i*) logically, a model trained on texts from a specific century works better on texts from that century, and vice versa (e.g. novels from the 19th century are more accurately analysed by the model trained on 19th century texts, than on the one trained on 18th or 20th century texts) and *ii*) more surprisingly maybe, the global model aggregating all the different sub-models works better than any other one on all the different corpora by a significant margin (*i.e.* by a statistically significant margin).

We can also observe that our results so far are not very impressive. There are several reasons for this, but the first one is clearly due to our approach. For both practical and theoretical reasons, our training sets are quite small, because we did not have enough time to provide large annotation sets and because we also wanted to avoid overfitting since we just considered a few works and a few authors (see above).

However, our results show that it is possible to develop only one model to annotate the different corpora, although each novel is specific. This is probably true because French has not evolved so much from the 18th century[1]. However, the model may still need some adaptation depending on the novels considered (some are known to have very specific ways to name people for example). This is why the ability to update an existing model and use active learning for training would be especially interesting in our case. It would also help to solve the annotation issue, since active learning makes it possible to reach high performance, while reducing drastically the annotation effort.

Lastly, we may want to explore neural network techniques for annotation, which are known to be slightly more efficient than CRF (Lample *et al.*, 2016), although, as said above, CRF as such are simpler and quite powerful for the annotation of continuous sequences.

6. Conclusion

We have presented an experiment aiming at showing that it is possible to develop accurate models for the annotation of characters in a corpus of French novels. Our current results, although far from perfect, are nevertheless sufficient for practical use. The next steps will consist in annotating many more novels and then develop large scale character analysis, i.e. detecting patterns in character networks, character interactions or character structures. For example, some novels put forward one character, or a few number of characters, whereas some others are based on the interaction of a larger group of characters. These

[1] This is also why we did not include older texts in our corpus: it is known that French has dramatically evolved in the 16th century, and even in the 17th, so it is advisable to be careful when dealing with texts prior to 1700 in French.

characteristics are known to be relevant for literary studies. The approach makes it possible to group together different kinds of novels, and also to check whether traditional classifications exhibit specific patterns, following some proposals made by Moretti (2005).

Acknowledgements

Thierry Poibeau is partially supported by the CNRS international Research Network "Cyclades". This work is supported by the cluster of excellence TransferS and by the Ecole universitaire de recherche (EUR) Translitterae (programme Investissements d'avenir ANR-10-IDEX-0001-02 PSL* et ANR-10-LABX-0099). Our thanks go to the anonymous reviewers for their constructive comments.

References

Dupont, Y. (2018). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. Conf. Traitement Automatique des Langues Naturelles (TALN), 2018.

Finkel, J.R.; Grenager, T.; and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 363-370.

Lafferty, J.; McCallum, A.; Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of the 18th International Conf. on Machine Learning,* Morgan Kaufmann, p. 282–289, 2001.

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. (2016). Neural architectures for named entity recognition. *Conf. of the North American Chapter of the Association for Computational Linguistics* (NAACL). San Diego, USA. p. 260-270.

Lavergne, T., Cappé, O. and Yvon, F. (2010). Practical Very Large Scale (CRFs). *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL), Uppsala, Sweden, p. 504—513.

Montani, I. and Honnibal M. (2017). Prodigy: A new annotation tool for radically efficient machine teaching. https://explosion.ai/blog/prodigy-annotation-tool-active-learning.

Moretti, F. (2005). Graphs, Maps, Trees. Abstract Models for a Literary History. London/New York, Verso, 2005, 160 p.

Moretti, F. (2013). *Distant Reading*. London/New-York, Verso, 244 p.

Piper, A.; Algee-Hewitt, M.; Sinha, K.; Ruths, D.; Vala, H. (2017). Studying Literary Characters and Character Networks. *Digital Humanities 2017*, Montreal, Canada.

Poibeau T., Kosseim L. (2001). Proper name extraction from non-journalistic texts. *Language and Computers* 37 (1), 144-157.

Poibeau T. (2003). The multilingual named entity recognition framework. *European Conference of the Association for Computational Linguistics* (EACL 2003).

Rosset, S., Grouin, C. and Zweigenbaum, P. (2011). Entités nommées structurées : guide d'annotation Quaero. Notes et Documents 2011-04, LIMSI, Orsay, France.

Benjamin Rabu
Frédérique Mélanie
Thierry Poibeau
Laboratoire LATTICE
(CNRS & ENS / PSL et U. Sorbonne nouvelle / USPC)
Ecole normale supérieure
1 rue Maurice Arnoux
92160 Montrouge, France
Mail: benjamin.rabu@icloud.com
Mail: frederique.melanie@ens.fr
Mail: thierry.poibeau}@ens.fr

Model	**F-measure per novel**					**Mean F-measure**
18th century model	De l’esprit des lois	Candide	L’an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.68**
	0.77	0.62	0.69	0.66	0.66	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.63**
		0.77	0.56	0.60	0.60	
			À l’ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.67**
			0.75	0.73	0.55	
19th century model	De l’esprit des lois	Candide	L’an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.61**
	0.55	0.65	0.69	0.63	0.53	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.70**
		0.76	0.65	0.73	0.68	
			À l’ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.67**
			0.70	0.74	0.57	

20th century model	De l’esprit des lois	Candide	L’an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.62**
	0.76	0.64	0.71	0.50	0.49	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.69**
		0.66	0.76	0.75	0.59	
			À l’ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.73**
			0.74	0.79	0.64	