Named Entity Recognition for French medieval charters

Sergio Torres Aguilar
École nationale des chartes
sergio.torres@chartes.psl.eu

Dominique Stutzmann
CNRS-IRHT
dominique.stutzmann@irht.cnrs.fr

Abstract

This paper presents the process of annotating and modelling a corpus to automatically detect named entities in medieval charters in French. It introduces a new annotated corpus and a new system which outperforms state-of-the art libraries. Charters are legal documents and among the most important historical sources for medieval studies as they reflect economic and social dynamics as well as the evolution of literacy and writing practices. Automatic detection of named entities greatly improves the access to these unstructured texts and facilitates historical research. The experiments described here are based on a corpus encompassing about 500k words (1200 charters) coming from three charter collections of the 13th and 14th centuries. We annotated the corpus and then trained two state-of-the art NLP libraries for Named Entity Recognition (Spacy and Flair) and a custom neural model (Bi-LSTM-CRF). The evaluation shows that all three models achieve a high performance rate on the test set and a high generalization capacity against two external corpora unseen during training. This paper describes the corpus and the annotation model, and discusses the issues related to the linguistic processing of medieval French and formulaic discourse, so as to interpret the results within a larger historical perspective.

1 Introduction

Named entity recognition (NER) is a fundamental task aiming at detecting and classifying words used as rigid designators in a text. Typically the operation consists in a lexical segmentation of texts separating entities from common words and a subsequent classification of them according to a set of predefined categories. NER has quickly become part of the NLP toolbox used by digital humanities to structure and mine textual collections. However, its application to historical texts still involves some challenges. Not only medieval charters use low-resource languages such as medieval versions of Latin until the 15th century and vernacular languages (e.g. Old and Middle French) from the 13th c. onwards, but they are also written in diverse linguistic versions due to language change over space and time. Moreover their strong topic-dependency further complicates the use of general classifiers and popular NLP tools used for modern languages, since charters contain mainly property deeds whose wording was framed by well-defined documentary models using stereotyped structures and a restricted and formulaic vocabulary.

In the last years, digitizing and scholarly editing original manuscript sources have gained great momentum, coinciding with their reappraisal in medieval studies. New tools are needed to explore and mine these specialized sources in order to foster insights from millions of documents and support new hypotheses. Named entity recognition, in particular, is a key element in order to provide an indexed structure to historical texts. It would allow the implementation of information retrieval techniques and adapt diplomatics and historical research methods to large scale corpora.

Our contribution can be summarized as follows: (1) An annotated corpus built upon three different collections of medieval acts in French, (2) an adequate training and validation framework, to create supervised models able to automatically distinguish places and person names in unstructured texts; (3) We suggest a test protocol to evaluate the models' ability to generalize on a wide range of acts regardless of regional and chronological differences.

2 Related work

Sequence tagging, including NER, is a classic NLP task. NER processing libraries such as Spacy, Flair, and Stanford CoreNLP have become popular in digital humanities, because one can easily train custom models and apply diverse neural approaches based on RNN architectures using control gates as in the case of LSTM approaches or adding a double-scope lecture (Bi-LSTM) (Schmitt et al.,

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 37–46
co-located with ICON-2021, December 16-19, 2021, Silchar, India.
https://doi.org/10.26615/978-952-94-5833-2_005

2019). Indeed, LSTM architectures using statistical classifiers such as CRF and LDA have become the most successful training methods, esp. when feature engineering cannot be deployed as it happens with low-resources languages. While supervised corpus-based methods have generally become the paradigm in NER research, NER approaches on ancient Western languages as classical Latin and Greek are still deploying ruled-based analyzers coupled with gazetteers and patronymic lists (Erdmann et al., 2016; Milanova et al., 2019) due to the lack of relevant annotated corpora. Moreover, research on NER for heritage resources focused on 19th century OCRed newspaper collections and it has not been confronted with the intense spelling variations of pre-modern, so-called "pre-orthographic" sources (Ehrmann et al., 2016; Kettunen et al., 2016).

Currently there are only few available NLP resources for pre-modern French: a dependency Treebank, the Syntactic Reference Corpus of Medieval French (Prévost and Stein, 2013), and two lemmatizers for Old French, one trained on the Base de Français Médiéval (Guillot et al., 2018) using TreeTagger and the other, Deucalion, based on a Encoder-Decoder architecture (Clérice and Camps, 2021). But there is a lack of large language models for tasks such as topic clustering and named entity recognition, given that PoS tools only detect, but do not classify proper names or deal with their length and composition.

(Zhang et al., 2020) show that robust NER models can be trained even in the absence of word-level features (e.g. lemma, POS), but that one then has to add character-based word representations to encode all lexical phenomena, concatenated with word embeddings vectors. Leveraging on pre-trained language models increase significantly the performance compared with traditional approaches (Ehrmann et al., 2021). Yet, static embeddings and contextualized word representations such as BERT (Devlin et al., 2018) and ELMO (Peters et al., 2018), require large-scale corpora for training and fine-tuning, and they are not available for ancient language versions ("*état de langue*") or domain-specific texts.

3 Corpus description

To remedy the lack of relevant training corpora, we created a relatively large dataset for the present task, composed of ca. 500,000 words, from three different sources: *Diplomata Belgica*, *HOME-Alcar*, and the *Arbois (CBMA)*.

3.1 Diplomata Belgica (DiBe)

The *Diplomata Belgica* [1] are a large database published by the Belgian Royal Historical Commission in 2014. It contains more than 35,000 critical references and almost 19,000 full transcriptions of mostly Latin and middle French charters (de Hemptinne et al., 2015). It is based on (Wauters and Halkin, 1866-1907; Bormans et al., 1907-1966). The edited charters range from the early 8th century (mostly royal diplomas) to the late 13th century with a high concentration on the final period from the mid-12th century (84% of the corpus). They are related to private and public business and issued by or for institutions and persons in nowadays Belgium and Northern France.

For this work, we have annotated 922 charters in French edited in the *Diplomata Belgica*. They all are dated in the 13th century, and transmit diverse legal actions (donations, privileges, concessions and confirmations, judicial sentences, sales and exchanges) concerning individuals and corporate bodies (lay or religious institutions). The main producers are: (1) aldermens (*échevins*) of Tournai, Arras and Cambrai, for 135 acts concerning mostly private business; (2) counts or countesses of Flanders, Hainaut, Laon, Bar, etc. for 164 acts, mostly for notifications and confirmations; (3) feudal lords for 214 acts, mostly donations and exchanges with abbeys, but also private business and charters of franchise for cities; (4) aldermen of Ypres, for 374 chirographs (i.e. charters produced in double or triple copy to give one to each stakeholder) written in the last third of the 13th century and concerning private affairs linked to trade and industry, e.g. sales, exchange contracts, loans, recognition of debts (Valeriola, 2019).

3.2 HOME-Alcar

The HOME-Alcar corpus published in 2021 (Stutzmann et al., 2021) was produced as part of the European research project *HOME History of Medieval Europe*, under the coordination of Institut de Recherche et d'Histoire des Textes (IRHT-CNRS). This corpus provides the images of medieval manuscripts aligned with their scholarly editions at line level as well as a complete annotation of named entities (persons and places), as a resource to train Handwritten Text Recognition

[1] https://www.diplomata-belgica.be/

	DiBe		Navarre		Fervaques		Saint_Denis		Arbois	
Acts (1190)	922		96		54		53		65	
Tokens (500 767)	311002		82878		24843		34275		47769	
category/ length	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC
1	3570 (38%)	8895 (89%)	534 (48%)	1834 (94%)	241 (65%)	589 (90%)	266 (42%)	498 (72%)	368 (33%)	1353 (95%)
2	2156 (23%)	789 (8%)	130 (12%)	70 (4%)	34 (9%)	41 (6%)	123 (19%)	119 (17%)	218 (20%)	25 (2%)
3	3254 (34%)	93 (1%)	379 (34%)	23 (1%)	74 (20%)	29 (4%)	206 (32%)	29 (4%)	397 (36%)	25 (2%)
>3	504 (5%)	225 (2%)	62 (6%)	27 (1%)	22 (6%)	2 (0.3%)	43 (7%)	45 (7%)	126 (11%)	18 (1%)
# entities	9484	10001	1105	1954	371	661	638	691	1109	1421
# tokens	19722	11670	2148	2138	635	766	1320	1011	2544	1549
Density	6.32 %	3.74 %	2.59 %	2.58 %	2.55 %	3.08 %	3.85 %	2.94 %	5.32 %	3.24 %
Normalized	9.12 %		4.72 %		4.75 %		6.29 %		7.66 %	

Table 1: Statistics on entities for each corpus according to their length. *Density* represents the percentage of tokens belonging to entities. *Normalized* expresses the sum of densities without taking in account the nested LOC cases, v.g. the locative in a person name.

(HTR) and NER models.

HOME-Alcar includes 17 cartularies, i.e. volumes containing copies of charters, produced by religious or public institutions to keep a memorial record of his properties and rights. These cartularies were produced between the 12th and 14th centuries. The corpus contains 3090 acts, with 2760 in Latin and 330 in Old and Middle French, and almost 1M tokens. Texts in French can be found in 12 of the 17 cartularies, but only in three they constitute a substantial part that is adequate to train NER models: (1) Cartulary of Charles II of Navarre : 96 acts (Lamazou-Duplan et al., 2010); (2) Cartulary of Fervaques abbey : 54 acts (Schabel and Friedman, 2020); so-called "White Cartulary" of Saint-Denis Abbey : 53 acts (Guyotjeannin, 2019).

The first one is from a lay family. The transcribed acts, dated between the 1297 and 1372, contain private donations and exchanges as well as other legal categories that are uncommon in religious cartularies, e.g. treatises, successions, indemnities. The other cartularies were produced by religious institutions, namely Norman and Ile-de-France abbeys respectively, and contain mostly donations from private persons and privileges from public authorities. The French acts are dated between 1250 and 1285 for Fervaques and between 1244 and 1300 for Saint-Denis.

3.3 Arbois (CBMA)

The cartulary of the city of Arbois in the Jura region was written in 1384, but largely keeps the language of the earlier originals, even of the 13th c. The edition contains 50 acts plus 32 acts in annex (65 in French, 17 in Latin) (Stouff, 1989). They are included in the Corpus Burgundiae Medii Aevi (CBMA, # 11424-11506)[2], which provides the base text for our annotation.

Arbois was part of the Burgundian county and obtained a charter of franchises in 1247. The community, though under the seigneurial regime, was recognised as a corporate entity and able to trade land and become an owner. The acts of the cartulary show its economical and social interactions with the lords or other communities: agreements about public issues such as military services and war costs, or about taxes and customs; charters declaring communal land purchases or lawsuits in court; even accounts that show the financial problems of the community due to the expenses of fortifications and wars. The community was ruled and represented by the aldermens (*prud'hommes*) who probably commissioned the redaction of the cartulary.

4 Corpus annotation

4.1 Annotation parameters

Since the early 11th century, personal names start adopting the name and by-name structure. They are composed by a baptism first name and a second part that can be a personal surname, a patronymic name (*nomen paternum*), or a locative. The latter form makes up to a third of all place entities and provides precious historical information as they typically correspond to micro-toponyms, whose existence is often not recorded otherwise. The annotation is focused on the proper name acting as a rigid designator and does not include co-occurrences as personal titles, dignities or functions. For example in the named entity expression: "Philippe, par la grace de dieu, roy de France" we annotate "Philippe" (PERS) and "France" (LOC), but not the full entity.

The annotation only records person and place names. Names of corporate bodies entities have

[2]http://www.cbma-project.eu/

TOKEN	PERS	LOC	TOKEN	PERS	LOC
Jehan	B-PERS	O	Vigilie	O	O
de	I-PERS	O	Nostre	O	O
Le	I-PERS	B-LOC	Dame	O	O
Capelle	I-PERS	I-LOC	Candeleir	O	O
Jehain	B-PERS	O	tous	O	O
chastelain	O	O	li	O	O
de	O	O	capitele	O	O
Cambray	O	B-LOC	de	O	O
et	O	O	Notre	O	B-LOC
seigneur	O	O	Dame	O	I-LOC
d'	O	O	de	O	I-LOC
Oisy	O	B-LOC	Cambray	O	I-LOC
Estienne	B-PERS	O	Margrite	B-PERS	O
Le	I-PERS	O	veve	O	O
Lonbart	I-PERS	O	Watier	B-PERS	O
Adan	B-PERS	O	sans	I-PERS	O
Bridoul	I-PERS	O	Paour	I-PERS	O
Huon	B-PERS	O	et	O	O
Le	I-PERS	O	Selie	B-PERS	O
Fevre	I-PERS	O	,	O	O
Jehan	B-PERS	O	se	O	O
Wilame	I-PERS	O	fille	O	O

Table 2: Example of annotations for named entities in DiBe # 15541, # 16356, # 17169, and # 36741

been annotated as organisations (ORG) in *Diplomata Belgica* first, but then folded to "places" (LOC) as in the other corpora, because these entities are mostly ambiguous in medieval texts (the church of "Notre Dame" or the lordship of "Oisy" mean a place and a corporate body at the same time).

4.2 Annotation process

The charters of the HOME-Alcar corpus were already annotated following a double scope: flat entities (proper names and simple periphrasis) and full entities (proper names and co-occurrences). This annotation was made on the basis of an automatic annotation using a multilingual NER model, then later corrected by two expert annotators. Inter-annotator agreement was not measured, as corrections and ambiguous cases were discussed among the annotators during the process.

The charters of *Diplomata Belgica* and Arbois charters were annotated in the flat style in the same manner. We first applied an automatic multilingual model and a single expert manually corrected the hypothesis.

We use the usual BIO format to encode the annotated labels as follows: B-tag, I-tag and O-tag to represent Begin (B) of label, continuation (I) of label and absence (O), respectively.

5 Training of the models

5.1 Data preparation

Our gold-standard (ground-truth) corpus is composed of 1190 acts ($\sim$ 0,5 M tokens), divided into two sets in order to conduct two experiments: (1) training and test on a homogeneous corpus; (2) test on additional, external corpora to measure the robustness of the model.

The first experiment is based on a corpus containing 1072 documents and encompassing the *Diplomata Belgica* and the cartularies of Navarre and Fervaques. It is randomly split with a 0.8-0.2 ratio: training set (844 documents), and validation and test sets (45 and 183 documents). The results of the first experiment are shown in Table 3.

The second experiment uses a corpus composed of two corpora unseen during the first experiment, the cartularies of Saint-Denis and Arbois (118 documents). The classifiers trained on the entire first corpus were applied on the second. The results are shown in tables 4 and 5.

For training we consider each charter as one training unit with a max length of 3,000 words (and a median of 276) and a max word length of 12 characters (and a median of 5).

5.2 Problem definition

We see our problem as a traditional two-step sequence labeling task. The input is a defined sequence of tokens $x = (x_1, x_2...x_{n-1}, x_n)$ and the output must be defined as a sequence of tokens labels $y = (y_1, y_2...y_{n-1}, y_n)$.

Both steps (PERS and LOC) may be combined, successive or separate. In our implementations, Flair and Spacy have independent annotation processes and our custom model has two successive steps.

5.3 The custom Bi-LSTM-CRF model

In the first step we extract word and sub-word features using NLP tools; the second step involves the training of the neural classifiers. For the custom Bi-LSTM model this step occurs in two stages. First, we apply the classifier to produce the places names hypotheses. Then, the classifier integrates the hypotheses of place names as an extra feature and predicts the person names.

5.3.1 Model Architecture

As is shown in Figure 1 we train three embedding vectors from our data. First is a word represen-

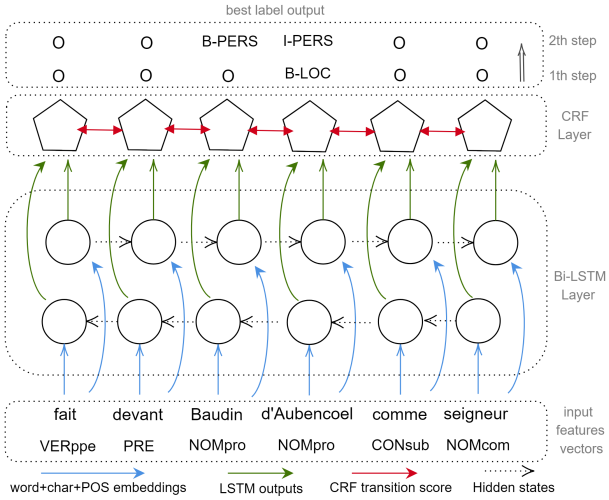

Figure 1: BiLSTM-CRF architecture using words, characters and POS embeddings as input on the excerpt "fait devant Baudin d'Aubencoel comme seigneur" (translates as: "written in front of Baudin d'Aubencoel acting as lord") tagged as a nested entity.

tation, second is a character-level representation, third is the POS character-level feature. Then, we merge these embeddings in order to form one single enriched vector and feed it into the Bi-LSTM. Finally the output hidden states are decoded using a CRF layer.

5.3.2 Character representations

Feeding the model with character-word information is a crucial step working with non-standard languages as it exploit sub-word level information as prefix and declension. In charters, many writing phenomena are linked to character variations due to the lack of grammatical rules or the introduction of spelling variations during redaction. In the same way, character-level representation naturally handles out-of-vocabulary as false lemmas, hapax and abbreviations, which can not be correctly expressed through word-vectors depending of a predefined dictionary and grammar pattern. Thus, character-level approaches are able to encode all textual phenomena at a morpheme-level using a restricted dictionary (115 keys in our work).

5.3.3 Word embeddings

Word embeddings improve the sequence tagging models. Yet, there are no publicly available embeddings for Middle French. In order to use pretrained embeddings we have trained a customized 300-dimensions and 44k vocabulary size word2vec model using a limited collections of medieval French charters (1.5M of tokens) which includes all the sub-corpora used in the present work plus French charters coming from other corpora as the CBMA and Île-de-France cartularies (Guyotjeannin, 2006-2010) [3]. Thus, the one-hot encoding words are replaced with their corresponding vectors.

5.3.4 POS information

The character-level embeddings do not catch morpho-syntactic information. To remedy this situation, POS features may be a workaround, specially working with large dependency plots as they import contextual features. But POS tags must be distributed among characters for each word. In this case inspired by the work of (Li et al., 2018), we generate a new feature combining character positions and POS tags. Positions are distributed using a 4-set tags as follows: B:Begin, M:Middle, E:end, S:single. The POS-tags were obtained using the TreeTagger lemmatizer made public by the Syntactic Reference Corpus of Medieval French in 2013 (Prévost and Stein, 2013).

5.3.5 Bi-LSTM-CRF Layer

Bidirectional long short-term memory (BiLSTM) models have proven to be effective for multiple sequence labelling tasks. As a classical RNN, the LSTM make output predictions based on long distance features using history information cells. The idea of the bidirectional variant is to reinforce the learning connecting the present and the past contexts of each token in the sentence. Thus, the output is a vector formed by the concatenation of a double sequence of LSTM hidden states $y_t = \overrightarrow{h_t} + \overleftarrow{h_t}$ for each token and token-features embeddings. This output is finally decoded by a Conditional Random Fields (CRF) layer which estimates the transition probabilities between tags and can predict the entire label sequence in each time step.

5.3.6 Training hyper-parameters

The grid search was evaluated on four key options: *batch-size* $\in$ { 2, 4, 16, 32 }, *output embeddings dimensions* $\in$ { 100, 200, 400 }, *learning methods* $\in$ { sgd, adam, rmsprop }, and *dropout* $\in$ { 0.2, 0.3, 0.4 } . Optimal combination was chosen following a 4-batch size, 200-dimensions embeddings, 0.2 dropout and rmsprop optimizer using a ReduceLROnPlateau scheduler.

[3] http://elec.enc.sorbonne.fr/

(a)

Model/ category	Flair			Spacy			Custom			Support
	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
B-PERS	0.938	0.959	0.949	0.948	0.969	0.958	0.975	0.982	**0.978**	2128
I-PERS	0.988	0.983	**0.986**	0.978	0.980	0.979	0.961	0.984	0.973	1805
micro avg	0.961	0.970	0.966	0.962	0.974	0.968	0.969	0.983	**0.977**	3933
B-LOC	0.960	0.945	0.952	0.956	0.952	0.954	0.967	0.975	**0.971**	2333
I-LOC	0.926	0.872	0.898	0.945	0.900	**0.922**	0.926	0.913	0.919	360
micro avg	0.956	0.935	0.945	0.954	0.945	0.950	0.961	0.966	**0.964**	2693

(b)

	PERS	LOC	PERS	LOC	PERS	LOC
Correct (TP)	2032	2187	2047	2213	2050	2210
Partial	16	27	26	24	47	33
Missed (FN)	80	119	55	96	31	90
Spurius (FP)	127	82	102	88	49	49
Pr	0.934	0.952	0.941	0.952	**0.955**	**0.964**
Rc	0.954	0.937	0.962	**0.948**	**0.963**	0.947
F1	0.944	0.945	0.951	0.950	**0.959**	**0.956**

Table 3: Evaluation results on test set for Flair, Spacy and Custom NER taggers : Pr (Precision), Rc (Recall), F1 (F1 score), TP (True positive), FN (False negative), FP (False positive), micro avg (micro-averaging score), Support (number of observations). First table (a) indicates tag-level performance; second table (b) indicates entity-level performance.

5.4 The SpaCy model

Spacy is an open-source NLP library proposing transformer-based pipelines. We fit our dataset on the default SpaCy NER architecture (Honnibal and Montani, 2017) which rely on two sequential encoding + attention networks: the first one is a typical embedding encoder which transforms tokens into a continuous vector space; the second one applies a manual extraction-features mechanism from the encoded tokens, evaluating the connections between tokens in a similar way of an attention layer. The goal is to give to each word a unique representation for each of its different contexts. Finally, a Multilayer Perceptron outputs the entity label. For the encoding step, we load in Spacy as backbone the transformers library CamemBert (Martin et al., 2019) trained on French modern texts.

For the Spacy model, as the default architecture does not accept other NER features, we trained two separate models for places and persons.

5.5 The Flair model

Flair is a PyTorch based NLP library which achieves state-of-art performance using pre-trained contextual embeddings as ELMO and BERT. The Flair NER architecture uses a deep learning architecture with Bi-LSTM layers in the back-end and allows users to activate CRF taggers (Akbik et al., 2018). In our case we deploy a special feature called "stacked embeddings" which combines classic embeddings with contextual embeddings in one single Pytorch vector. In our modelling we used the FastText (Bojanowski et al., 2017) embeddings trained on French Wikipedia as word-vectors combined to the default Flair multilingual forward + backward contextual embeddings. As for Spacy, we trained two separate Flair models for places and persons.

6 Evaluation

Table 3 shows the best results obtained with a training set of 1 072 charters. We provide the usual Precision, Recall and F1-score metrics at a token-level (B- and I- tags). We also include full-entity level metrics on strict match: strict match occurs when the hypothesis and the ground-truth match perfectly.

All three models obtain high performance results, both in PERS (0.944 to 0.959) and LOC (0.945 to 0.956) categories. The first metric shows that performance between B- and I- tags are harmonic which implies that our three models are able to correctly detect the boundaries of the entities regardless of their length. The second metric confirms that false negatives and false positives are marginal both in LOC and PERS thus achieving a very good result in multi-class tasks.

The custom Bi-LSTM shows a better performance in most categories. Specifically it generalizes better as indicated by the lower number of false negatives and false positives compared to the Flair and Spacy models. It can be explained by the use of a denser set of features including PoS and French medieval embeddings more adapted to the medieval charters. However, the Flair and Spacy performance is only 1 to 2 points lower, confirming

Saint-Denis

	Model/	Flair			Spacy			Custom			Support
	category	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
	B-PERS	0.959	0.964	0.962	0.954	0.970	0.962	0.982	0.969	**0.975**	638
	I-PERS	0.955	0.930	0.942	0.945	0.959	0.952	0.975	0.961	**0.968**	682
(a)	micro avg	0.957	0.946	0.952	0.962	0.974	0.968	0.978	0.975	**0.976**	1320
	B-LOC	0.898	0.916	0.907	0.905	0.926	0.915	0.926	0.942	**0.934**	691
	I-LOC	0.885	0.863	0.873	0.899	0.855	0.877	0.904	0.873	**0.889**	320
	micro avg	0.894	0.899	0.896	0.903	0.903	0.903	0.919	0.920	**0.920**	1011
		PERS	LOC		PERS	LOC		PERS	LOC		
	Correct (TP)	596	627		598	633		597	638		
	Partial	19	29		21	29		27	36		
(b)	Missed (FN)	23	35		19	29		14	17		
	Spurius (FP)	26	49		29	41		9	38		
	Pr	0.930	0.890		0.922	**0.900**		**0.943**	0.896		
	Rc	0.934	0.907		**0.937**	0.916		0.936	**0.923**		
	F1	0.932	0.898		0.930	0.908		**0.939**	**0.909**		

Table 4: Results of model evaluation on the cartulary of Saint Denis.

Arbois

	Model/	Flair			Spacy			Custom			Support
	category	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
	B-PERS	0.945	0.920	0.932	0.933	0.944	0.938	0.922	0.953	**0.938**	1109
	I-PERS	0.940	0.954	**0.947**	0.942	0.944	0.943	0.966	0.890	0.927	1435
(a)	micro avg	0.942	0.939	0.941	0.938	0.944	**0.941**	0.947	0.917	0.932	2544
	B-LOC	0.936	0.939	0.938	0.927	0.953	0.940	0.923	0.968	**0.945**	1421
	I-LOC	0.746	0.758	0.752	0.773	0.791	0.782	0.837	0.798	**0.817**	128
	micro avg	0.920	0.924	0.922	0.915	0.939	0.927	0.916	0.954	**0.934**	1549
		PERS	LOC		PERS	LOC		PERS	LOC		
	Correct (TP)	982	1326		998	1350		979	1362		
	Partial	50	28		57	18		95	20		
(b)	Missed (FN)	77	67		54	53		35	39		
	Spurius (FP)	47	73		62	91		76	108		
	Pr	**0.910**	**0.929**		0.893	0.925		0.851	0.914		
	Rc	0.885	0.933		**0.900**	0.950		0.883	**0.958**		
	F1	**0.897**	0.931		0.897	**0.937**		0.867	0.936		

Table 5: Results of model evaluation on the cartulary of Arbois.

that a robust model for the recognition of nested entities can be trained on our corpus using domain unspecific embeddings or without additional linguistic features, which are not usually available for ancient language versions.

In general, the three models perform slightly less well on I-LOC suggesting some issues on LOC compound entities. Average results are not affected severely since only 5-10% of the entities belongs to this type (see table 1). A close inspection reveals that I-LOC errors are propagated due to the presence of some uncommon and large entities (*Saint Nicolay des Pres en costé Tournay*, *Jehan Godin de Maisieres sur Meuse*) in which the models make a partial or a double annotation. In other cases, errors are triggered for LOC or PERS by the presence of large periphrastic denominations (*Margherite suer Abreham Bertelot veve Jehan le Crudenere*) or for less common entity particles as the article "li" (*Jehans li Cornus del Fontenil*) or the Flemish preposition *van* (*Watier van Wormezeele*). In some other cases, particles associated to regular vocabulary as *des*, *d'*, *de la*, *dou* which can be connected to multiple contexts, may induce a partial match (*Willaumes dou Temple li Macheclers*, *Saint Johan du Pié des Pors*).

6.1 Evaluation on external corpora

Tables 4 and 5 show that the results obtained on the test set are largely replicated when the models are applied to external corpora (Arbois and Saint Denis), with a harmonic precision and recall for the three models, a similar performance in B- and I-tags and a high performance in strict match ranging from 0.897 to 0.939 in PERS and from 0.898 to 0.937 in LOC, thus confirming that generalization in external documents only causes a small drop in performance (2 to 3 points). Again the custom

Bi-LSTM model seems to perform better than the Flair and Spacy models.

In general, charters from Saint-Denis are closer in style to the ones in *Diplomata Belgica* and Fervaques because they record similar legal actions and uses similar formulas. The performance is slightly lower in Arbois which is a municipal cartulary with a later chronology registering many juridical actions rarely found in the training corpus.

As in the test set, the main problem remains the recognition on the I-LOC tag. The same kind of errors appears as in the first experiment, but most are triggered by festivity names, a type that had not been annotated. In charters, dates are indicated using the saints' festivities and appear using a saint's name which was learned as a LOC entity by the model (*a paier le ior Saint Martin*, *chascun an es huiteves Saint Denys*). In consequence in some cases the models may propose a false positive.

7 Discussion

This work clearly proves that robust tools to classify named entities on French medieval charters can be modeled using character-based neural network approaches (Bi-LSTM-CRF) on our annotated corpus. The test sets being of different dates (end of 13th c. and end of 14th c.), the high performance proves that the models are fairly robust against language change. This may be explained as follows.

1) In Northern France, during the late Middle Ages, the anthroponymic structure is stable. Not only the stock of first names is limited (e.g. in *Diplomata Belgica*, 47% of persons use one of the ten top names *Jehans, Watiers, Jakemes, Willaumes, Henris, Pieres, Nicholes, Bauduins, Gilles, Margherite*, and their variants), but also by-name and even the periphrastic denomination follow a recurrent pattern in which the model easily fits.

2) In a similar way, the named entity co-occurrences, which are crucial to calculate transition scores, belongs to a restricted stock. In charters there are a constant reference to a well-delineated territorial space as well as to broad system of titles, offices and dignities presenting a person. For example, in *Diplomata Belgica*, five terms (*sire/messire, bourgois, signor/monsigneur, eschevin, dame/madame*) co-occur in 24% of all personal entities.

3) Moreover, the lexical and semantic contexts of appearance of the named entities are well-defined by the use of formulaic models. Formulas are not fixed and charters are not mass-produced nor standardized, but they involve the use of a restricted vocabulary and are constrained by their need to follow a certain form, since they are documents with legal value.

These circumstances greatly help to obtain a valid NER model starting from a limited collection of charters. We have demonstrated that custom and off-the-shelf library models are able to capture the underlying structure of the charters' entities even using a small set of features and can be successfully applied to other diplomatic collections in spite of chronological and regional differences. Most errors concern partial matches on untypical data or complex data on which the model fits hardly because certain lexical series are missed or hindered.

8 Conclusion

We present an annotated corpus to French medieval charters and three neural NER models. The evaluation returns a strong performance reaching 0.96 in both PERS and LOC categories on the homogeneous test set and 0.95 in PERS and 0.92 in LOC on unseen data which confirms that the model can be used on charters from other chronologies and origins.

Besides, we can confirm that our models are able to produce a double hypothesis which implies a high confidence on the recognition of nested entities extensively used in medieval charters.

While this work concerns the development of a neural NER model, it can benefit several research areas including indexation systems, data visualization and distant reading methods. Named entities are the base of several digital and classical humanities research methods on networks, timelines, event-lines and GIS-maps. These models and the annotated data on which it is built, which are themselves new contributions, can be easily integrated into other pipelines, thus contributing to enhance the toolbox for Old and Middle French regarding other supervised methods.

9 Model repositories

The models, source code and the annotated corpora supporting this work are available at (Torres Aguilar, 2021)

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Stanislas Bormans, Fabienne Marien, Joseph Halkin, J. Cuvelier, Jean-Jacques Hoebanx, and Charles Wirtz. 1907-1966. *Table chronologique des chartes et diplômes imprimés concernant l'histoire de la Belgique*. Commission royale d'histoire, Palais des Académies, Bruxelles.

Thibault Clérice and Jean-Baptiste Camps. 2021. chartes/deucalion-model-af: 0.4.0Alpha.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, CONF, pages 97–107. Bochumer Linguistische Arbeitsberichte.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*.

Alex Erdmann, Christopher Brown, Brian D Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for latin named entity recognition. In *COLING 2016: 26th International Conference on Computational Linguistics*, pages 85–93. Association for Computational Linguistics.

Céline Guillot, Serge Heiden, and Alexei Lavrentiev. 2018. Base de français médiéval: une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, (7):168–184.

Olivier Guyotjeannin. 2006-2010. Cartulaires numérisés d'ile-de-france.

Olivier Guyotjeannin. 2019. Chartes de l'abbaye de Saint-Denis.

Thérèse de Hemptinne, Jeroen Deploige, Jean-Louis Kupper, and Walter Prevenier. 2015. Diplomata belgica: les sources diplomatiques des pays-bas méridionaux au moyen âge. the diplomatic sources from the medieval southern low countries.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2016. Old content and modern tools-searching named entities in a finnish ocred historical newspaper collection 1771-1910. *arXiv preprint arXiv:1611.02839*.

Véronique Lamazou-Duplan, Anne Goulet, and Philippe Charon. 2010. *Le cartulaire dit de Charles Ii roi de Navarre*. Presses universitaires de Pau et des Pays de l'Adour.

Yanzeng Li, Tingwen Liu, Diying Li, Quangang Li, Jinqiao Shi, and Yanqiu Wang. 2018. Character-based bilstm-crf incorporating pos and dictionaries for chinese opinion target extraction. In *Asian Conference on Machine Learning*, pages 518–533.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Ivona Milanova, Jurij Silc, Miha Serucnik, Tome Eftimov, and Hristijan Gjoreski. 2019. Locale: A rule-based location named-entity recognition method for latin text. In *HistoInformatics@ TPDL*, pages 13–20.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Sophie Prévost and Achim Stein. 2013. Syntactic reference corpus of medieval french (srcmf). *Lyon & Stuttgart: ENS de Lyon*.

Chris Schabel and Russell L. Friedman. 2020. *The Cartulary of Fervaques Abbey, a Cistercian Nunnery*. in press.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343. IEEE.

Louis Stouff. 1989. *Cartulaire de la ville d'Arbois au comté de Bourgogne*. Revue bourguignonne de l'enseignement supérieur, 8, n° 2.

Dominique Stutzmann, Sergio Torres Aguilar, and Paul Chaffenet. 2021. HOME-Alcar: Aligned and Annotated Cartularies. Type: dataset.

Sergio Torres Aguilar. 2021. Named Entity Recognition for French medieval charters. Models and datasets.

Sébastien de Valeriola. 2019. Le corpus des chirographes yprois, témoin essentiel d'un réseau de crédit du xiiie siècle. *Bulletin de la Commission royale d'Histoire*, 185(1):5–74.

Alphonse Wauters and J. Halkin. 1866-1907. *Table chronologique des chartes et diplômes imprimés concernant l'histoire de la Belgique*. M. Hayez, Bruxelles.

Yu Zhang, Zhenghua Li, Houquan Zhou, and Min Zhang. 2020. Is pos tagging necessary or even helpful for neural dependency parsing?