Multilingual Named Entity Recognition for Medieval Charters using Stacked Embeddings and BERT-based Models

Sergio Torres Aguilar

Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 119–128
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Multilingual Named Entity Recognition for Medieval Charters using Stacked Embeddings and BERT-based Models

Sergio Torres Aguilar
École nationale des chartes, Centre Jean-Mabillon, Paris, France
sergio.torres@chartes.psl.eu

Abstract

In recent years the availability of medieval charter texts has increased thanks to advances in OCR and HTR techniques. But the lack of models that automatically structure the textual output continues to hinder the extraction of large-scale lectures from these historical sources that are among the most important for medieval studies. This paper presents the process of annotating and modelling a corpus to automatically detect named entities in medieval charters in Latin, French and Spanish and address the problem of multilingual writing practices in the Late Middle Ages. It introduces a new annotated multilingual corpus and presents a training pipeline using two approaches: (1) a method using contextual and static embeddings coupled to a Bi-LSTM-CRF classifier; (2) a fine-tuning method using the pre-trained multilingual BERT and RoBERTa models. The experiments described here are based on a corpus encompassing about 2.3M words (7576 charters) coming from five charter collections ranging from the 10th to the 15th centuries. The evaluation proves that both multilingual classifiers based on general purpose models and those specifically designed achieve high-performance results and do not show performance drop compared to their monolingual counterparts. This paper describes the corpus and the annotation guideline, and discusses the issues related to the linguistic of the charters, the multilingual writing practices, so as to interpret the results within a larger historical perspective.

Keywords: Latin NER, old spanish NER, old french NER, medieval NLP, NLP for historical languages

1. Introduction

Named entity recognition (NER) is one of the first steps towards information extraction aiming at locating words used as rigid designators in an unstructured text and classify them according to a set of predefined categories such as person names, locations and organizations. NER has quickly become part of the Natural Language Processing (NLP) toolboxes used to structuring and mining vast textual collections. However, its application to ancient and pre-orthographic texts still involves some challenges. In the case of medieval charters, we can mention the following:

Low-resources language varieties : Medieval charters are written in medieval versions of Latin until the 15th century and vernacular languages (e.g Old and middle French, old Spanish) from the 13th c. onwards. Annotated corpora for these languages are still rare preventing the developing of powerful and adapted NLP toolboxes. In addition to this, the written testimonies show different language states defined by more or less important linguistic changes over time and space which complicates generalization model capacities.

Multilingualism : Multilingual NER libraries are quite recent and the overall performance is usually lower compared to the monolingual systems. Charter collections dated from the mid-13th century display documents in both Latin and vernacular languages. Public powers continue to use *scripta latina*, especially for solemn documents, until the end of the Middle Ages; while vernacularization of private documents occurs since the late 12th century (Glessgen, 2004). Code-switching practices and bilingual sequences can be detected even within the same charter, as in the case of the *vidimus* : a charter for revalidating old rights that includes a verbatim copy of the original act issued in Latin; or in the case of the late use by notaries of long-established Latin formulae in the legal language of the acts. (See two bilingual charters in the annexus).

Strong topic-dependency : Charters are legal deeds whose wording was framed by well-defined documentary models using stereotyped discursive structures and a formulaic and archaizing vocabulary. Charters are not mass productions, but they use a series of more or less recurrent sequences according to their typologies and the legal actions recorded in the document. This stands for a fundamental problem when using popular classifiers since they hardly fit on this kind of documents whose syntax and semantics may be largely unknown to an out-of-the-box classifier trained on present-day discourse from news and Wikipedia.

Complex denomination : Nested entities and context ambiguity are open questions in modern NER research. Most of the NER classifiers work in a flat mode while in medieval texts, nested entities are quite common in the form of locatives, patronyms and periphrasis coupled with baptismal names as a strategy of social distinction against a high homonym ratio. On the other hand, the concept of moral person, common category in modern NER works, is relatively foreign in charters, since most organizations are presented in an ambiguous manner using the context of locations from which it is often very difficult to distinguish them.

These four aspects of charters will be explored in our experiments in the aim of creating robust multilingual named entities models to provide an indexed structure to historical collections that can potentially con-

tain texts with nested entities as well as different languages and language states. These efficient NER models would allow the implementation of information retrieval techniques and adapt diplomatic and historical research methods to large scale corpora.

Our contribution can be summarized as follows: (1) An annotated multilingual corpus built upon five different collections of medieval charters in a range of five centuries (10th to 15th), (2) an adequate training and validation framework, to create supervised NER models able to automatically distinguish places and person names in unstructured multilingual texts; (3) A robustness protocol to evaluate the models' ability to generalize on a wide range of acts regardless of regional, typological and chronological differences.

2. Related work

NER is a classic sequence classification task. Traditionally the best neural approaches for NER were based on LSTM or Bi-LSTM approaches working with word and character-level representations. Lately, these approaches were partially replaced by the transformers architectures based on attention mechanisms as they eliminate the vanishing gradient problem providing direct connections between the encoder states and the decoder. Recently the use of pre-trained contextual language representations such as BERT (Devlin et al., 2018) and ELMO (Peters et al., 2018) have become the standard for sequence classification as they can be fine-tune on many downstream tasks in a supervised fashion.

The leveraging on these pre-trained models increases significantly the performance compared with traditional word-based approaches (Ehrmann et al., 2021) and eliminates the need to deploy methods depending on rich features engineering in favor of fine-tuning processing based on the update of word and sub-word representations from labeled data. Yet, contextualized word representations and even static embeddings require large-scale annotated corpora for training and fine-tuning, and their adaptation to ancient language versions («*états de langue*») or domain-specific texts has not been fully studied. An advanced version of these models such as mBERT (Devlin et al., 2018) and RoBERTa (Conneau et al., 2019) trained on multilingual big datasets has proven that it is possible to generalize across languages and get powerful models capable of handling tasks in a multilingual environment.

Despite this, some popular NER systems on ancient Western languages are still deploying ruled-based analyzers coupled with gazetteers and patronymic lists (Erdmann et al., 2016; McDonough et al., 2019) due to the lack of relevant annotated corpora which block the deploying of supervised approaches, while others are skill-dependent using the NLP tools for ancient languages that have been published in the last years. Indeed, some lemmatization and PoS tools are available for ancient languages (Clérice et al., 2019; Prévost and Stein, 2013). But there is a lack of large language models for tasks such as text classification and NER, given that PoS tools only detect, but do not classify proper names or deal with their length and composition. And in the best of our knowledge any NLP resource exists to treat medieval documents at a multilingual level.

3. Corpus description

To remedy the lack of relevant training corpora, we created a relatively large dataset for the present task, composed of ca. 2.3 millions of tokens, from four database sources ranging from 10th to the 15th century (See figure 1): *Diplomata Belgica* (de Hemptinne et al., 2015), *HOME-Alcar* (Stutzmann et al., 2021), the *CBMA* (Magnani, 2020) and the *Codea* (Borja, 2012) corpus. The first three contain Latin and French charters while the CODEA corpus concentrates on old Spanish. Furthermore, we have annotated two other single cartularies, taken here as external datasets, for testing the classifier robustness: the cartularies of the seigneury of Nesle (1217-1282) (Hélary, 2007) and of the monastery of Eslonza (912-1399) (Vignau, 1885) written in French-Latin and Spanish-Latin respectively.

3.1. The CBMA

The CBMA (*Corpus de la Bourgogne du Moyen Âge)* is a large database composed of about 29k charters coming from the Burgundy region dated between the 9th and 14th centuries. Since 2016 the CBMA project has made freely available a sub-corpus of 5300 manually annotated charters with named entities. This subcorpus constitutes the core component of our modeling for medieval Latin. The documents it contains, coming from nearly a hundred small localities in Burgundy, are taken from ten different cartularies, i.e, volumes containing copies of charters about land exchanges, public privileges concessions, disputes, contracts, papal letters, etc. The preparation of these volumes was normally undertaken by religious or public institutions with the aim of keeping a memorial record of their history but also to serve as a source of legal proofs about rights and properties acquired by donation or purchase. Most part of annotated CBMA documents are in Latin coming from private persons and public institutions. Many French charters can be found in the corpus but they were not originally included on the annotated subset. To extend the annotations for French, we have selected and annotated the cartulary of the city of Arbois (Stouff, 1989) belonging to the same collection. This is a municipal cartulary commissioned in 1384 by the aldermens (*prud'hommes*) of the city and contains documentary types that can hardly be found in the cartularies from religious institutions : 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, reflecting the economic and social interactions between the community and the lords or other communities.

	LATIN		FRENCH		SPANISH	
Acts (7576)	5474		1245		857	
Tokens (2.3M)	1.36M		0.53M		0.51M	
CBMA	5282		65		-	
DIBE	-		922		-	
HOME	39		203		-	
CODEA	77		-		800	
Nesle	28		55		-	
Arlanza	48		-		57	
category/ length	PERS	LOC	PERS	LOC	PERS	LOC
1	66921 (91%)	33291 (71%)	6079 (42%)	14391 (90%)	5381 (34%)	15610 (89%)
2	3173 (4%)	9841 (21%)	2849 (19%)	1057 (7%)	7998 (50%)	883 (5%)
3	3178 (5%)	986 (2%)	4703 (33%)	245 (1%)	1068 (7%)	364 (2%)
>3	743 (1%)	2607 (6%)	812 (6%)	356 (2%)	1490 (9%)	558 (3%)
# entities	74015	46735	14443	16049	15937	17415
# tokens	85976	69435	29348	18739	30823	20855
Density	6.31%	5.10%	5.51%	3.51%	6.09%	4.12%
Flat Density	11.08%		8.17%		9.73%	

Table 1: Statistics on entities for each corpus according to their length (number of tokens). *Density* represents the percentage of tokens in the whole corpus annotated as entities. *Flat density* expresses the sum of densities without taking in account the nested LOC cases, v.g. the locative in a person name.

3.2. The Diplomata Belgica (DiBe)

The *Diplomata Belgica* are a large database published by the Belgian Royal Historical Commission in 2014. It contains almost 19,000 full transcriptions of mostly Latin and middle French charters. It is based on (Wauters and Halkin, 1866 1907; Bormans et al., 1907 1966). The edited charters range from the early 8th century to the late 13th century with a high concentration on the 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 all the French charters (922 docs) 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). In this sub-corpus are also included 374 chirographs (i.e. charters produced in double or triple copy to give one to each stakeholder) from the aldermen of Ypres, concerning private affairs linked to trade and industry, e.g. sales, exchange contracts, loans, recognition of debts (Valeriola, 2019).

3.3. HOME-Alcar

The HOME-Alcar corpus (Stutzmann et al., 2021) was produced as part of the European research project *HOME History of Medieval Europe*. This corpus provides the images of medieval manuscripts aligned with their scholarly editions as well as an annotation of named entities (persons and places), in the aim to serve as a resource to train synchronously Handwritten Text Recognition (HTR) and NER models.

HOME-Alcar includes 17 cartularies dated between the 12th and 14th centuries. The corpus has 3090 acts (2760 in Latin, 330 in Old and Middle French) and almost 1M tokens. From this corpus we have selected French charters coming from four cartularies: (1) Cartulary of Charles II of Navarre : 96 acts (Lamazou-Duplan et al., 2010); (2) the Cartulary of seigneury of Nesle; 83 acts (Hélary, 2007); (3) Cartulary of Fervaques abbey : 54 acts (Schabel and Friedman, 2020); (4) the so-called «White Cartulary» of Saint-Denis Abbey : 53 acts (Guyotjeannin, 2019)

The first two are from lay families. In the case of Navarre, 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. In the case of the cartulary of Nesle compiled in the 1270s, it contains documents related to purchases, debts, distribution of inheritances, land disputes, which attempt to accurately describe the patrimony of Jean, lord of Nesle. The other two were produced by religious institutions, namely Norman and Ile-de-France abbeys respectively, and have mostly donations from lay people and privileges from public authorities. The French acts are dated between 1250 and 1285 for Fervaques and between 1244 and 1300 for Saint-Denis.

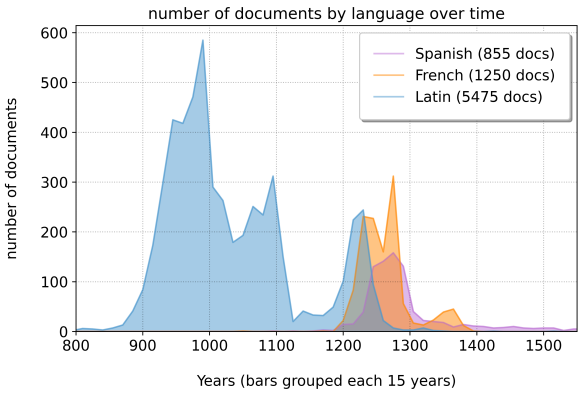

Figure 1: Number of documents over the time by languages including all the 7.6k documents

3.4. The CODEA corpus

The CODEA corpus (*Corpus of Spanish Documents Prior to 1800*) is a free available corpus made public in 2012 by the University of Alcalá. Its main objective is tracing the evolution of Spanish from the High Middle Ages to the emergence of modern Spanish (Borja, 2012). The origins of the documents are quite diverse as the CODEA team tries to generate a plural image that includes charters from different regions of the Iberian Peninsula (but mainly from the former Aragon and Castille areas), as well as from different social states and institutions : chancelleries and city offices, but also notaries and small scriptoria. These diachronic series aim to facilitate the analysis of changes in the written language and in writing practices considering the social, economic and institutional origin of the document. Consequently, the typological variety of CODEA charters is quite wide, since we have chancery documents: privileges, mandates, provisions, grants; private charters as contracts, sales, letters, wills, and normative documents : regulations, reports, inventories. Unlike the aforementioned collections, the number of charters coming from ecclesiastical institutions is small including these from the papacy that continues to write in Latin until after the Middle Ages.

Today the corpus contains 2,500 charters, ranging from the 11th to 16th centuries. To enrich our model with Spanish named entities, we have chosen and annotated a random sub-set of 877 documents of which we can say that 800 are written in Spanish, or mostly in Spanish, and 77 in Latin, or mostly in Latin, since clear linguistic separations are in some cases impossible.

3.5. The Eslonza cartulary

The scholarly edition of the cartulary of Eslonza was published in 1895 (Vignau, 1885), it contains the charters transcriptions from the cartulary of the Benedictine monastery of San Pedro de Eslonza (León, Spain) founded in 1099. The cartulary contains 227 acts (57 in medieval Spanish) dated between the 912 and the 1350. As in other cartularies from religious institutions the acts are related to land exchanges and business between the abbey and public and private persons. Some acts are dated prior to the foundation of the abbey and some other describe exchanges between two lay landowners. This is explained because when a monastery inherited a land from lay people the charters attesting the legal origin of this land were also transferred and preserved as legal guarantee. These documents defined as *munimina* by diplomatics appear together with *instrumenta*, solemn acts such as diplomas where the monastery is the author and recipient of an act that attests the receipt of a property or a right, later validated by an authority.

4. Corpus annotation

4.1. Annotation parameters

Our annotation is focused on the named entities considered as rigid designators including proper names and excluding pronouns, co-occurrences terms and complex periphrasis, which form the so-called «full-entity», because they contain words belonging to the dictionary. For example in the case of the full-entity « *don Suero Pérez , obispo de Çamora* » we annotate « *Suero Pérez* » (PERS) and « *Çamora* » (LOC) but we do not include the honorific prefix: « *don* » (Lord, dominus) and the dignity title: « *obispo* » (bishop).

In addition, we annotate the nested entities which are detected in charters since the early 11th century and whose use became the norm since the late-12th century. The composition of these nested entities, also called «by-names», varies according to the regions and times, but in general the structure is composed by either a locative or a patronym (*nomen paternum*) or both coupled to a baptism name by declension of using a nexus. These added locatives provide precious historical information as they typically correspond to microtoponyms, whose existence is often not recorded otherwise. In these cases, a «LOC» tag is partially aligned to a «PERS» entity. For example : *Matheus Guidonis d'Attrebato; Bartolome de Moral del Payuelo*.

Furthermore, our annotation only records person and place names. The corporate bodies entities, normally annotated as organizations (ORG), where folded to «places» (LOC) as in the other corpora, because they are mostly ambiguous in medieval texts. In many cases a same entity can be a reference to an institution, a building or a land: the cathedral of «Saint-Vincent» or the lordship of «Oisy» mean a place and a corporate body at the same time. In other cases, it is unclear if a name involved in an action refers to a land, a corporal body or a moral person: as for example: a land donation to «Sanctus Petrus» is made materially to a monastery, but under the patronage of the saint to whom it is dedicated. The annotation of ORG entities needs the use of external resources for disambiguation and a LOC tag must be preferred for these cases.

4.2. Annotation process

The charters of the HOME-Alcar corpus were already annotated and corrected by two experts following a double scope: single 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 CRF-NER model, then later corrected by two expert annotators. The alpha Inter-annotator agreement was not measured.

The charters of *Diplomata Belgica*, Arbois, CODEA and Arlanza were annotated in the single-entity style in the same manner. A single expert manually corrected an automatic first 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. During the robustness test we also add a special «L(location)-PERS» tag to mark nested location entities in a flat-mode.

Token	Nested		Flat	Flat-nested
Magister	O	O	O	O
Iobertus	B-PERS	O	B-PERS	B-PERS
de	I-PERS	O	I-PERS	I-PERS
Ponte	I-PERS	B-LOC	I-PERS	L-PERS
curie	O	O	O	O
Senonensis	O	B-LOC	B-LOC	B-LOC
officialis	O	O	O	O
don	O	O	O	O
Pedro	B-PERS	O	B-PERS	B-PERS
de	I-PERS	O	I-PERS	I-PERS
Leorna	I-PERS	B-LOC	I-PERS	L-PERS
abat	O	O	O	O
del	O	O	O	O
monasterio	O	O	O	O
de	O	O	O	O
Santa	O	B-LOC	B-LOC	B-LOC
María	O	I-LOC	I-LOC	I-LOC
de	O	I-LOC	I-LOC	I-LOC
Valbuena	O	I-LOC	I-LOC	I-LOC

Table 2: Example of annotations for named entities in # CBMA 18296 , and # CODEA-0346

5. Training of the models

5.1. Data preparation

Our ground-truth corpus is composed of 7576 acts ($\sim$ 2,31 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 7388 acts (177253 annotated entities) and encompassing the charters from the five aforementioned corpora. It is randomly split with a 0.8 - 0.2 ratio: training set (5911 acts), and validation and test sets (441 and 1036 acts). This experiment consists of two steps: in the first, we train three monolingual models (table 3); in the second, we train multilingual models in order to compare performances (table 4). The second experiment tests the generalization capacity of the models on a unseen corpus : the cartularies of Eslonza (105 charters) containing Latin and Spanish charters and Nesle (83 charters) containing Latin and French charters. The monolingual and multilingual classifiers trained on the entire first corpus were applied on the second (table 6).

5.2. Problem definition

We see our problem as a traditional 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 $y = (y_1, y_2...y_{n-1}, y_n)$. We have implemented three training modes following the nested nature of the entities: The first one operates in a nested mode and both steps (PERS and LOC) have independent training processes using two classifiers; the second operates in a Flat (multi-class) mode, that means PERS and LOC are recognized in a synchronous manner without overlapping; the third, introduces a «L-PERS» special tag (see table 2) in the traditional BIO-format with the aim of recognizing cases of nested entities (locatives within personal names) using a single classifier. Results are presented in table 5.

5.3. The BERT-based models : mBERT and XLM-RoBERTa

We fine-tune two multilingual BERT varieties: on the one hand, mBERT (Devlin et al., 2018) which uses a 12 multi-head attention layers like the BERT-Base model but instead of being trained on raw English texts it is trained on Wikipedia pages of 104 languages. On the other hand, XLM-RoBERTa (Conneau et al., 2019) which is a large model using 24 layers and trained on 10 times more data that mBERT.

Both BERT-based encoders learn on a massive amount of raw data a deep language representation in an unsupervised way generating an embedding contextualized vector for each input token. In contrast to classic sequence models that predict the next word, BERT tries to optimize a Masked Language Model (MLM) objective and next-sentences prediction thus performing contextual token encoding and understanding the relationship between two contiguous sequences. XLM-RoBERTa optimizes the same MLM objective but prefers a dynamic masking during the training.

In these approaches the training is done in a unsupervised fashion without any alignment between the languages. Instead of using specific language vocabularies they introduce a shared vocabulary which activate cross-lingual transfer operations during training and fine-tuning. This reduces complexity of space and helps the model learn the underlying structure of a language rather than just learning the monolingual vocabulary. Several experiments prove that both mBERT and RoBERTa perform well in cross-lingual generalization for a variety of downstream tasks. (Muller et al., 2020; Conneau et al., 2019)

Training a NER BERT-based classifier is a three-steps task: Firstly, we vectorize the sentence and label sequences using the BERT-based word-pieces tokenizer; secondly, we freeze all the layers except the last in order to keep the pre-trained weights; finally, we pass the annotated data through the final layer, thus partially retrained the model using a cross-entropy loss function.

5.3.1. Hyperparameters

We fine-tune the models to perform sub-word level classification over sentences with a max-length of 250 word-pieces. Each model was fine-tuned over 5 epochs starting in a 2.0e-5 learning rate. We ran a 16 batch and AdamW as dynamic optimizer. In addition, since BERT models relies on word-pieces tokenizations (v.g: "Garner", "##us", "Dei", "gra", "##tia", "Tre", ##cens", "epi", "##sco", "##pus") which do not match the original token-split annotation, we decide, as was done in the original BERT paper, to train the model on the tag labels for the first word piece token of each word.

5.4. The stacked embeddings model (+Bi-LSTM-CRF)

Bi-directional LSTM classifiers using a final CRF-layer are one of the most used architectures for addressing sequence tagging tasks. Used together with static

Lang /	Latin				French				Spanish			
Category	Pr	Rc	F1	Support	Pr	Rc	F1	Support	Pr	Rc	F1	Support
B-PERS	99.2	99.0	99.1	10052	96.6	97.5	97.0	1811	99.1	98.9	99.0	2248
I-PERS	95.8	94.0	94.9	1808	97.8	98.7	98.3	1973	99.6	98.3	98.9	1911
micro avg	98.7	98.2	**98.5**	11860	97.2	98.2	**97.7**	3784	99.3	98.6	**99.0**	4159
B-LOC	97.2	98.1	97.6	6204	97.5	96.5	97.0	2493	98.7	99.2	99.0	2605
I-LOC	96.7	95.7	96.2	2848	92.1	93.6	92.8	359	93.1	96.7	94.9	306
micro avg	97.0	97.3	**97.2**	9052	96.8	96.1	**96.5**	2852	98.1	99.0	**98.5**	2911

Table 3: Evaluation results on test set for the monolingual models using the bi-LSTM-CRF + stacked embeddings architecture : Pr (Precision), Rc (Recall), F1 (F1 score), micro avg (micro-averaging score), Support (number of observations).

(a)

Model /	Combined			Multi_Flair			Multi_BERT			XLM_RoBERTa			
Category	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Support
B-PERS	98.9	98.8	98.8	98.8	98.9	98.8	98.3	98.9	98.6	98.9	98.8	**98.9**	14111
I-PERS	97.8	97.1	97.4	97.9	96.4	97.1	97.8	96.8	97.3	97.6	97.4	**97.5**	5692
micro avg	98.6	98.3	98.4	98.6	98.2	98.4	98.2	98.3	98.2	98.5	98.4	**98.5**	19803
B-LOC	97.6	98.0	97.8	97.6	98.5	**98.0**	97.6	97.7	97.7	97.8	97.7	97.8	11302
I-LOC	95.9	95.6	95.7	96.1	96.3	**96.2**	94.5	95.8	95.2	95.0	96.2	95.6	3513
micro avg	97.2	97.4	97.3	97.3	98.0	**97.6**	96.8	97.3	97.1	97.1	97.4	97.2	14815

(b)

	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC
Correct (TP)	13839	11012	13841	11071	13848	10990	13842	10988
Incorrect	126	124	133	111	133	141	132	147
Missed (FN)	146	166	137	120	130	171	137	167
Spurius (FP)	137	213	137	219	223	213	142	174
Pr	98.1	97.0	98.0	97.1	97.5	96.8	98.0	97.2
Rc	98.0	97.4	98.1	98.0	98.1	97.2	98.1	97.2
F1	98.1	97.2	**98.1**	**97.5**	97.8	97.0	98.1	97.2

Table 4: Evaluation results on test set for the multilingual models using the bi-LSTM-CRF + stacked embeddings architecture (Combined and Multi_Flair) and the fine-tuned BERT-based models (mBERT and XLM-RoBERTa) : TP (True positive), FN (False negative), FP (False positive). First table (a) indicates tag-level performance; second table (b) indicates entity-level performance.

embeddings and later with contextual embeddings became popular in recent years for NER tasks. Recent works indicate that stacking both classes of embeddings by concatenating and remapping them can significantly improve performance (Catelli et al., 2020), especially in multilingual environments when the pretraining languages have similar characteristics (Akbik et al., 2018). We think that the stacking strategy can also be effective when working with old linguistic varieties, since words and sub-words embeddings can help to deal with both the polysemy of the language and the inconsistency in spelling.

We train the Bi-LSTM-CRF classifier using Flair, one of the state-of-the-art Library in NLP tasks, based on Pytorch and natively supporting the stacking of embeddings. The contextual embeddings were trained on a concatenated trilingual corpus of 20M of words from medieval charters using the contextual embeddings Flair model that capture latent syntactic-semantic information (Akbik et al., 2018). The static embeddings were generated from this same corpus using Fasttext word-representation which is trained on subword-level information (Bojanowski et al., 2017).

5.4.1. Hyperparameters

The FastText embeddings were training with 200 dimensions using a skipgram model. The Flair embeddings were training in a bidirectional mode using a 1024 hidden-size and a maximum sequence length of 250 tokens. As for the Bi-LSTM classifier, the grid search was evaluated on three key options: batch-size {4,16,32}, starting learning rate {1.0e-2, 2.0e-2, 5.0e-3} and hidden size {256, 512}.

6. Evaluation

Table 3 shows the best results obtained for the three monolingual classifiers using the Bi-LSTM-CRF + stacked embeddings architecture. 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. These models were trained by choosing the charters that correspond to each language within the train, test and dev datasets (see 4.1). Table 4 shows the best results for the multilingual models using the adapted and the fine-tuned BERT methods. These models were trained on the entire datasets. The «combined» column of table 4 concatenates the inferences of the 3 monolingual models in order to compare performances of the lingual-specialized models against the cross-lingual models as they are trained and tested on the same data.

Summarizing over the results we can state that multilingual models (table 4) do not show a performance loss over their monolingual counterparts. Except in the case of I-LOC, we can state that the differences between all the models are marginal. But we must emphasize that while the multilingual models use just two classifiers (PERS and LOC) the monolingual ones use 6 (PERS and LOC x 3 languages) to reach the same result.

	Flat-nested mode									Flat mode									
	Multi_Flair			Multi_BERT			XLM-RoBERTa			Multi_Flair			Multi_BERT			XLM-RoBERTa			
	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Support
B-PERS	99.0	98.8	**98.9**	98.6	98.9	98.7	98.7	98.9	98.8	99.0	99.0	**99.0**	97.4	97.2	97.3	98.9	98.8	98.9	14111
I-PERS	97.9	97.0	**97.4**	97.6	96.8	97.2	97.2	96.9	97.1	98.1	97.2	**97.7**	95.4	97.4	96.4	97.7	97.3	97.5	4148 / 5692
L-PERS	97.7	97.0	**97.3**	96.2	96.6	96.4	96.5	96.8	96.6	-	-	-	-	-	-	-	-	-	1544 / 0
B-LOC	97.2	98.1	**97.7**	97.2	97.4	97.3	97.1	97.7	97.4	97.1	98.2	**97.7**	94.8	94.6	94.7	97.1	97.8	97.4	9877
I-LOC	95.6	96.1	**95.9**	94.2	95.9	95.0	94.2	95.8	95.0	95.4	96.3	**95.9**	92.3	90.5	91.4	94.2	96.3	95.2	3394
micro avg	97.9	98.0	**98.0**	97.5	97.8	97.6	97.4	97.9	97.7	97.9	98.2	**98.1**	95.9	96.1	96.0	97.6	98.0	97.8	33074

Table 5: Evaluation results on test set for the multilingual Flat and Flat-nested models using the bi-LSTM-CRF + stacked embeddings architecture and the fine-tuned BERT-based models.

	Eslonza + Nesle											
Model /	Combined			Multi_Flair			Multi_BERT			XLM-RoBERTa		
category	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Support
B-PERS	97.7	97.9	97.8	98.3	97.9	**98.1**	96.4	96.8	96.6	97.9	97.1	97.5	4697
I-PERS	97.0	97.6	97.3	97.4	97.2	**97.3**	95.8	95.9	95.9	97.0	96.7	96.8	3302
micro avg	97.4	97.8	97.6	98.0	97.6	**97.8**	96.1	96.4	96.3	97.5	96.9	97.2	7999
B-LOC	96.4	95.0	95.7	96.8	96.3	**96.5**	93.3	95.4	94.3	95.7	95.0	95.3	2896
I-LOC	95.7	89.8	92.6	95.2	92.5	**93.8**	92.6	87.9	90.2	95.2	89.5	92.3	1104
micro avg	96.2	93.5	94.9	96.4	95.2	**95.8**	93.1	93.3	93.2	95.6	93.4	94.5	4000
	PERS	LOC		PERS	LOC		PERS	LOC		PERS	LOC		
Correct (TP)	4537	2715		4542	2749		4490	2705		4484	2698		
Incorrect	97	54		88	58		100	97		117	91		
Missed (FN)	63	127		67	89		107	94		96	107		
Spurius (FP)	74	84		49	73		143	175		78	99		
Pr	96.3	95.1		97.0	95.4		94.9	90.9		95.8	93.4		
Rc	96.6	93.7		96.7	95.0		95.5	93.4		95.5	93.2		
F1	96.4	94.4		**96.9**	**95.2**		95.2	92.0		95.6	93.3		

Table 6: Evaluation results on external test set for the multilingual models using the bi-LSTM-CRF + stacked embeddings architecture and the fine-tuned BERT-based models.

Both monolingual and multilingual models obtain high performance results in PERS (98% in average) and LOC (97% in average) categories also showing an harmonic recall and precision along the categories. As is often seen in NER for ancient texts, the detection of I-classes is slightly lower due to ambiguities and imbalances in the corpus, since complex (multi-word) entities, especially in places, represent a low percentage of the total (see table 1). But in general, we realize that all the models can correctly detect the boundaries of the entities regardless of their length.

Furthermore, the mBERT and especially RoBERTa models achieve almost the same result as Flair-based ones by fine-tuning a general-purpose model without formally requiring external embeddings, which are generally not available for historical texts. Thus, demonstrating that an adaptation of BERT during 5 epochs (2 hours in a RTX3090) could be enough to obtain a suitable model for applications to medieval texts.

Switching to a Flat mode (table 5) does not mean an improvement in performance. Training in this mode may seem easier than a nested mode, but in the latter, there is actually a smaller number of categories to classify. Although the task seems much more complicated for BERT who is outperformed by RoBERTa and Flair.

In the same way, the Flat-nested model (table 5) shows an almost identical result to the Flat mode on a task that is slightly more complex, thus proving that a single classifier can be enough to obtain an excellent result (98%), just below the best performance (98.1%), in a multilingual, multi-class and multi-label NER task.

6.1. Evaluation on external corpora

Table 6 shows the predictions of the multilingual and monolingual models on the external test corpus (Eslonza + Nesle). The proportion of shared entity mentions between these corpus and the training corpus is of the order of 27% for personal names (mostly common baptismal names) and 25% for place names.

Again, we can state a very high precision and recall in the recognition of the personal name (97% in exact-match) and slightly lower on locations (95% in exact-match). The drop in performance with respect to the test corpus is quite low (1 to 2 points), thus confirming that all our classifiers have an acceptable generalization capacity on unknown documents.

On the other hand, the Flair model is more competent when facing unseen documents than BERT and RoBERTa, who present a much higher number of false negatives and false positives. Analyzing much more closely his inferences we can detect two kind of errors: label misclassification (v.g *Alffonso Martines alcalde del Rey* (true: B-PERS ; false: B-LOC); *Pero Breton* (true: I-PERS ; false: B-LOC) *archipreste*) and confusion between NEs and non-NEs classes (v.g : *in octabis Sancti Martini* (true: O-O ; false: B-LOC-I-LOC); *in festo beati Andree* (true: O ; false: B-PERS) *Apostoli*). The first ones correspond to contextual errors whose presence does not seem aberrant; while the second ones correspond to errors about the dates (not annotated in our corpus), since in the Middle Ages, they were written down using saints' festivities, who are sometimes recognized by the classifiers as location entities.

7. Discussion

This work clearly proves that high-performance NER classifiers for medieval charters can be modeled using neural approaches and pre-trained models. The results on the test sets being multilingual and multi-regional proves the models are robust against changes on personal denomination traditions, chronologies and language. This may be explained as follows:

1) Although personal naming strategies change according to regional traditions, they follow a recurring pattern in which the model easily fits. As we have seen, the so-called by-name does not stand for an insurmountable issue for the models that captures well the periphrases and couplings used in the formation of the by-name and is even capable of classifying the nested location entity. This has a lot to do with the fact that the stock of personal names is relatively restricted. In the case of *CODEA* charters with a chronology limited to two centuries (13th-14th) the 52% of persons take one of the top ten names: *Pedro, Fernando, Joán, Alfonso, Martín, Domingo, Sancho, Rodrigo, García, María*. While in a more diverse corpus with a longer chronology (10th-14th) like the CBMA the concentration is less dense, but still very high compared to modern standards as at least 18% of people take one of the top ten names : *Hugues, Bernardus, Rotbertus, Petrus, Stephanus, Durannus, Willelmus, Iohannes, Odon, Arnulfus*, and their variants (v.g for *Rotbertus* : *Rotbertus, Rodberto, Robertus, Rotbert*).

2) In an analogous way, the entity co-occurrences, which are crucial to calculate transition scores in a contextual way, belongs to a restricted stock. The charters use a stable and shared vocabulary that reflects the social and administrative order and tries to specify the legal action as much as possible. A broad but regular system of titles, dignities and offices is activated to specify the category of benefactors, recipients and witnesses and usually precede the personal name. For example, in *Diplomata Belgica*, five terms (*sire/messire, bourgois, signor/monsigneur, eschevin, dame/madame*) cooccur in 24% of all personal entities. Similarly, space is well delineated both in the consciousness of men and in scriptural practices, and a hierarchical order of territorial organization serves as a coordinate system to spatially locate the movable and immovable property that is the object of the exchange. This can be verified in the CBMA corpus where the six top spatial words: *uilla, pagus, terra, ecclesia, locum, ager* cooccurs with almost a third (32%) of the total locations entities since it is the most common vocabulary, before the 13th century, for spatial determination in land transfers, the most abundant legal action in this corpus.

3) The formulaic nature of the charters proposes a relatively stable discursive structure. The documents follow a model according to the type of act and the legal action and are individualized by particular information such as the named entities. The charters have parts that support a freer wording, for example those dedicated to explaining the background and the conditions of the exchange and other more constrained ones, such as naming the authors and witnesses, the dates and the validation signs. This structure facilitates the identification of the entities since it reduces the complexity of the sequences and the probability distribution for the predictions. Certainly, the charters are not mass-produced and most of the formulas are not strictly fixed, but during their drafting, a restricted vocabulary and a regulated discursive form are used, since this is one of the elements that give the charter its value as legal proof.

4) Moreover, acts with legal value follow similar writing forms throughout Western Europe based on Latin legal language and Latin formularies. The change from the Latin code to the vernacular languages does not occur drastically and supposes the coexistence of documents of similar value written in both languages over a period of several centuries. The scribes continue to use common Latin formulation, a legal vocabulary set by prestige and tradition and following a discursive format typical of the Latin legal act in their intention to communicate a legal action clearly and explicitly. Phenomena such as linguistic interference, bilingualism, literal translations and code-switching are common in late medieval written production. Languages codes appear to be interchangeable or specific to certain situations and form a «charter language» with high semantic and lexical overlapping between Latin and vernacular languages. These overlaps favor the cross-lingual generalization during modeling since the shared words and structures are mapped onto similar representations at the same time as their co-occurrences, thus spreading the generalization effect over other word pieces of the sequence. These circumstances greatly help to create multilingual models whose performances are competitive with their monolingual counterparts.

8. Conclusion

We present an annotated multilingual corpus of medieval charters to address NER tasks. We have demonstrated that fine-tuned on general-purpose models and off-the-shelf library architectures are able to capture the underlying structure of the charters' entities in a multilingual environment reaching an average of 98% in the recognition of persons and places names. Our evaluation on unseen data confirms that they can be successfully applied to other diplomatic collections despite chronological, regional and linguistic differences. Besides, we can confirm that our models are able to produce a multi-class hypothesis using a single classifier which implies a high confidence on the recognition of nested entities extensively used in medieval charters. These models and the annotated data on which they are built, which are themselves new contributions, can be easily integrated into other pipelines, thus contributing to enhance the toolbox for the automatic treatment of the medieval text regarding other supervised methods and other Latin-derived languages.

9. Model repositories

The models, source code and the annotated corpora supporting this work are available at (Torres Aguilar, 2022) and at our git repository: `https://gitlab.com/magistermilitum//ner_medieval_multilingual/`

10. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Borja, P. S.-P. (2012). Desarrollo y explotación del «corpus de documentos españoles anteriores a 1700»(codea). *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, (1):5–35.

Bormans, S., Marien, F., Halkin, J., Cuvelier, J., Hoebanx, J.-J., and Wirtz, C. (1907-1966). *Table chronologique des chartes et diplômes imprimés concernant l'histoire de la Belgique*. Commission royale d'histoire, Palais des Académies, Bruxelles.

Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., and Esposito, M. (2020). Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Applied Soft Computing*, 97:106779.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

de Hemptinne, T., Deploige, J., Kupper, J.-L., and Prevenier, W. (2015). *Diplomata Belgica: les sources diplomatiques des Pays-Bas méridionaux au Moyen Âge.* Commission royale d'Histoire.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A. (2021). Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*.

Erdmann, A., Brown, C., Joseph, B. D., Janse, M., Ajaka, P., Elsner, M., and de Marneffe, M.-C. (2016). Challenges and solutions for latin named entity recognition. In *COLING 2016: 26th International Conference on Computational Linguistics*, pages 85–93. ACL.

Glessgen, M.-D. (2004). L'écrit documentaire dans l'histoire linguistique de la france. *La langue des actes. Actes du XIe Congrès international de diplomatique*.

Guyotjeannin, O. (2019). Édition électronique des chartes de l'abbaye de Saint-Denis, 2019.

Hélary, X. (2007). L'édition électronique du cartulaire de la seigneurie de nesle. *Bulletin du centre d'études médiévales d'Auxerre| BUCEMA*, (11).

Lamazou-Duplan, V., Goulet, A., and Charon, P. (2010). *Le cartulaire dit de Charles Ii roi de Navarre*. Presses universitaires de Pau et des Pays de l'Adour.

Magnani, E. (2020). Des chartae au corpus: la plateforme des cbma-chartae/corpus burgundiae medii aevi. *Digitizing Medieval Sources. Challenges and Methodologies.*, pages 57–67.

McDonough, K., Moncla, L., and Van de Camp, M. (2019). Named entity recognition goes to old regime france: geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, 33(12):2498–2522.

Muller, B., Anastasopoulos, A., Sagot, B., and Seddah, D. (2020). When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Schabel, C. and Friedman, R. L. (2020). *The Cartulary of Fervaques Abbey, a Cistercian Nunnery*. in press.

Stouff, L. (1989). *Cartulaire de la ville d'Arbois au comté de Bourgogne*. Revue bourguignonne de l'enseignement supérieur, 8, nº 2.

Stutzmann, D., Torres Aguilar, S., and Chaffenet, P. (2021). HOME-Alcar: Aligned and Annotated Cartularies. Zenodo: https://doi.org/10.5281/zenodo.5600884.

Valeriola, S. d. (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.

Vignau, V. (1885). *Cartulario del Monasterio de Eslonza*. Madrid : Imp. de la viuda de Hernando y Cª.

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

11. Language Resource References

Clérice, Thibault and Camps, Jean-Baptiste and Pinche, Ariane. (2019). *Deucalion, Modèle Ancien Francais (0.2.0)*. Zenodo. https://doi.org/10.5281/zenodo.3237455.

Prévost, Sophie and Stein, Achim. (2013). *Syntactic reference corpus of Medieval French (SRCMF)*.

Torres Aguilar, Sergio. (2022). *Multilingual named entity recognition for medieval charters. Datasets and models*. Zenodo. https://doi.org/10.5281/zenodo.6463699.

12. Appendix

Two annotated examples (red for persons, blue for places) of bilingual charters in the CBMA (Charter 1) and in the CODEA (Charter 2) : Common Latin formulations in the charter protocols and the notarized act in vernacular.

Charter 1. **CBMA 18859**. Vidimus (1273), by Hugues, archdeacon of Langres, of the charter according to which Jehanz, parish priest of Châteauvillain and dean of Chaumont, makes it known that Renauz li Acuers d'Orges admitted having donated to the abbey of Vauxbons one piece of land located at village of Orges (1256). AD Haute-Marne, 1 H 84, pièce no 8. Chauvin Benoît, L'abbaye de moniales cisterciennes de Vauxbons au diocèse de Langres (… 1175 - 1394…). Étude historique et édition du chartrier, Devecey, 2004, A27.

Universis presentes litteras inspecturis, Hugo, archidiaconus Lingonensis, salutem in Domino. Noveritis quod nos litteras inferius annotatas vidimus et verbo ad verbum legimus non cancellatas, non abolitas nec in aliqua sui parte viciatas quarum tenor talis est : Nos Jehanz, curez de Chastelvilen et doien de Chaumont, fasonz savoir a tout cas qui verrunt ces presantes latres que an ma presance estaubliz Renauz li Acuers d'Orges qui fuit filz Roelim la Lemont d'Orges a requeneu qu'il a donei de dei et d'armone por l'ame de ces acesors un jornel de terre[1] a l'abaiassa de Valbaion et es dames de leaus, li quez jornez de terre siet or finaige d'Orges ce est a savoir es seillons dares la maison au palletz aupres [2] Jaquel le Graure Chapusot. Au tamonnaige de laquel chosse, a la requeste de l'une partie et de l'autre, nos avons mis notre seel en ces presantes latres. Ce fut fait an l'an de grace mil et IIC et LVI, or mois de mars lou vanredi davant lou diemange que on chante Letare Jherusalem[3]. In cuius rei testimonium presenti transscripti, sigillum nostrum apposuimus. Datum a nobis die sabbati ante festum beate Marie Magdalenes anno Domini M°CC° septuagesimo tercio, mense julio.

Charter 2. **CODEA 0231**. Charter of sale (1216) by which Ordón Pédrez de Cavia sells several real estate properties (lands, wastelands, meadows, etc.) in the locality of Cillamayor (Palencia) to Taresa Verbúdez for 50 maravedies. Archivo Histórico Nacional, Clero, Palencia, carpeta 1653, nº 16.

In Dei nomine et eius gratia. Notum sit omnibus hominibus tam presentibus quam futuris quod ego Ordon Pedrez de Cavia vendo illa hereditate quantam habeo en Cellamayor e en Alfoz de Santo Juliano e illa renta que habeo en santa Juliana de Candiola : solares, los poblados e los ermos, plados e tierras, et illa parte de la eclesia de Santa Maria de Cellamayor, esto es, la cuarta parte quod fuit de donna Urraca Ferrandez mea avola, vendo a donna Taresa Verbudez por L morabetis et sum pacati de precio e de rovla[4]. Et si aliquis homo de mea progenie vel de extranea istam cartam voluerit disrumpere sit ille maledictus e excomunicatus cum Judas traditore in inferno dampnatus et pecet in coto[5] C morabetis regi terre. Facta carta in era MCCLIIII[6] regnante rege don Anric in Toleto e in Castella. Alferaz el conde don Alvaro, mayordomus don Gonzalvo Roiz, merino mayor Ordon Martinez, episcopus en Burgos maestre Mauriz. Hec sunt testes estantes e videndentes: Gonzalvo Garciaz de Grajera, Gonzalvo Johanes de Quintana Tello, Alvar Munioz de Rebiela [...] Petrus Isidorus qui notuit.

[1] A *journal* is a unit of land measurement (corresponds approximately to the area worked by a man in a day).
[2] Translated as: «[This land is located] in the furrows behind the fence of the house of Jaquel le Graure Chapusot».
[3] The «Laetare Ierusalem» was sung the fourth Sunday in the season of Lent (Laetare Sunday).
[4] Translated as: «The price and the *robra* are agreed». The *robra* was a treat paid by the buyer to close a sale.
[5] *Pechar in coto* is a common sanction formula that orders the person who opposes the contract to pay a sum as compensation.
[6] The calculation of the data according to the Hispanic Era starts from the year 38 BC.