DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 4

A Named Entity Recognition Model for Medieval Latin Charters

Pierre Chastang <pierre_dot_chastang_at_uvsq_dot_fr>, UVSQ-Université Paris-Saclay
Sergio Torres Aguilar <regester3_at_gmail_dot_com>, UVSQ-Université Paris-Saclay
Xavier Tannier <xavier_dot_tannier_at_sorbonne-universite_dot_fr>, Sorbonne Université

Abstract

Named entity recognition is an advantageous technique with an increasing presence in digital humanities. In theory, automatic detection and recovery of named entities can provide new ways of looking up unedited information in edited sources and can allow the parsing of a massive amount of data in a short time for supporting historical hypotheses. In this paper, we detail the implementation of a model for automatic named entity recognition in medieval Latin sources and we test its robustness on different datasets. Different models were trained on a vast dataset of Burgundian diplomatic charters from the 9th to 14th centuries and validated by using general and century *ad hoc* models tested on short sets of Parisian, English, Italian and Spanish charters. We present the results of cross-validation in each case and we discuss the implications of these results for the history of medieval place-names and personal names.

1. Introduction

In this paper, we introduce a new model for the automatic recognition of named entities in medieval Latin charters. The named entity concept involves all physical and real objects which can be designated with a specific name. A named entity recognition (NER) model detects entities in the text and classifies them (e.g., a person's name, a location name, an organization's name, a date). It becomes possible, then, to index the resulting information units and to extract relationships between them. In that sense, automatic NER systems can help to improve current information systems by building a knowledge base of all the spatial, social and personal entities in a set of documents; this enables a searchable and quotable corpus matrix using actors and places as dynamic keywords which help to provide meaningful answers to questions asked on a large textual database. 1

In recent years, some projects have entailed the production of structured data through manual tagging of named entities in thousand-item historical corpora. This manual operation takes years for an entire team and can obviously not be scaled up to the almost 500,000 (and increasing) medieval digitized documents. This lengthy process can be completed only with massive acceleration through a computational operation of converting automatically thousands of documents in raw data format into structured data in a short time. 2

In the present project we use one of these hand-annotated corpora as our dataset: the CBMA (*Corpus Burgundiae Medii Aevi*), which contains a collection of medieval charters and cartularies produced in Burgundy. From the 11th century on, many ecclesiastical institutions started to compose cartularies in which they transcribed their own documents, especially titles relating to their property and their land rights, in order to better preserve written deeds and keep legal proof of ecclesiastical possessions and social relationships. Today, these registers stand as one of the major sources for medieval studies. Among the texts collected in the CBMA we isolated a set of 5300-items (1,2 million words), which correspond mostly to cartularies from Cluny. This large sub-set contains mainly donations, exchanges and sales, that is, written records dealing with the transfer of lands, goods and rights between ecclesiastical institutions and between private and public actors. While these documents are relatively formulaic, with only limited vocabulary and a relatively small number of turns of phrase, their geographical provenance is as diverse (more than one hundred different places) as the chronological scope of their production is wide (from the early 10th century to the middle of the 3

13th century). Moreover, they include many different kinds of documents: charters, notices, bulls, diplomas, letters, lawsuits, etc.

As we will show, developing a model able to achieve entity recognition for this corpus presents several challenges. The need to choose a single corpus for stability and coherence can affect the ability of the model to be generalized to a wide range of documents, because it reduces the variety of samples and can hide more specific phenomena. Constructing a robust model requires finding a technical and intellectual compromise between collecting sufficiently varied data and avoiding extreme training on some similar structures [Plank 2016]. This issue becomes even more complicated if we consider the particular characteristics of our main corpora: on the one hand, diplomatic charters combine formulaic structure and variable lists of named entities [Zimmermann 2003], which can be an advantage regarding iteration in training. But it also means fitting problems of recognition over other corpora, because there are many formulaic traditions, different models in according to juridical actions and regional redactional dependencies. On the other hand, because Latin is a low-resource language in natural language processing, the work must be conducted using tools and vocabularies not initially developed for Latin or developed for classical Latin literature [Eger et al. 2015]; not to mention the technical adaptations required for the specificities of medieval Latin due to several linguistics alterations. Above all, less technical but more extended, pre-processing must be completed, such as the normalization of textual spelling, the manual validation of much inflected language arbitrariness, and the listing of difficulties due to overlapping and to textual ambiguities. 4

The construction of a proper validation environment, proving the robustness of our model, has led us to build several corpora and sub-corpora variations and to annotate manually other smaller corpora. We finally developed a validation protocol with an extensive set of evaluations in order to show that our model can be applied to a wide range of unannotated charters with different typologies and belonging to different scriptural traditions and regional origins. 5

This paper describes, first, the historical nature of our corpus and the features of our gold-standard corpus. It explains the method adopted in generating the model with a machine-learning approach based on conditional random fields (CRF). We detail the construction of the model and sub-models, as well as the results obtained by the application of our general model and by the different experiments using cross-validation. In the discussion, we try to develop a broader argument for results by referring to current onomastic studies, historical and scriptural usages, and language variations, thus offering a contextualized explanation that can help to clarify, from a non-technical and humanistic perspective, matches and errors obtained in the application of the model. The constitutive models of this project have been deployed in a web-based application adapted to expert and non-expert users located at https://entities-recognition.irht.cnrs.fr. An automatic and accessible web-based annotation workflow, based on models developed in this project, has been deployed at https://entities-recognition.irht.cnrs.fr. Finally, we propose some suggestions for the fruitful application of these models in the wider context of digital humanities. 6

2. Academic Background

Named entity recognition (NER) is one of the most promising technical tools in the digital humanities field. Some classical digital humanities projects have been conducted with NER, thereby opening new ways to explore and query large databases. Actively used in the medical, biological and journalistic domains [Abacha et al. 2011] [Eltyeb et al. 2014], named entities prove to be meaningful pieces of information intimately connected to the main issues addressed by humanistic studies. NER techniques are an excellent way to provide an overview of a corpus. On the one hand, they help to classify data into pre-defined categories, which is primordial to indexing and describing data and eventually to building structured databases. On the other hand, they can activate data exploration at different scope levels. For example, NER techniques can be used to identify core concepts and cluster them into vocabularies, ontologies, and actor's attributes on documentary series They are also indispensable in the pre-treatment of texts since they provide pieces of information that do not appear in the language dictionaries. Alternatively, at a corpora level, NER techniques can serve to build thick social, spatial and semantic relationship networks using named entities as nodes. 7

NER work in the digital humanities community consists of two subtasks: (i) detection and classification of named entities in classical and modern language texts and (ii) entity linking towards already existing knowledge databases. Concerning 8

the first task, experiments have proven that supervised approaches are more suitable for large-scale databases than rule-based or unsupervised ones [Nadeau et al. 2007]. In these supervised methods, especially statistics-sequential methods, algorithms use labelled data and meta-information obtained by different types of syntactical analysis, such as POS tagging, parsing, chunking, from word-based to sentence-based levels. The system then builds a statistical model based on these labelled data and suggests the best labels for a new, unlabelled sequence of words. Some digital humanities projects in recent years [Curran et al. 2003] [Won et al. 2018] [Rayson et al. 2017] produce supervised models exceeding 80% on the F-measure (harmonic mean of precision and recall).

Nevertheless, because of the lack of annotated data, rule-based or dictionary-based models are still very popular, despite the fact that they generally achieve a low recall ratio (for the former) or a low precision ratio (for the latter). Training a dictionary-based recognition model against a list of names can lead to a high ratio of recognition for a particular corpus, but the model is often not robust when applied to unseen texts or different types of data. On the other hand, rule-based models trained using rule definitions and descriptions to categorize entities show a valid global recall, but a slight tendency to poor precision-ratio on unseen texts. In spite of that, these methods have proven very useful for rapid textual annotation of small corpora or more standardized corpora like those from literature and journals [Grover et al. 2008] [Mosallam et al. 2014] [Ehrmann et al. 2016]. 9

Recent years have seen an increasing popularity of hybrid, semi-supervised approaches. Bootstrapping techniques use a small set of annotated data in order to obtain automatically tagged data by taking advantage of specific repetitive and contextual patterns. Some work on this topic has introduced interesting propositions with excellent results [Brooke et al. 2016] [Neelakantan et al. 2015] [Klein et al. 2014]. 10

Both unsupervised and supervised methods need annotated corpora in order to build better linguistic analyzers and evaluation tools. This problem has motivated, in recent years, different actions in the digital humanities community. The main one is the production of handcrafted corpora with one or more fully annotated attributes, such as morphological features, semantical descriptions, syntactic relationships, named entities, etc. The medieval corpora made available by CBMA, CDLM, and SRCMF projects are among these, as well as examples of well-constructed lists of authorities, gazetteers and treebanks such as proposed by Pelagios, LASLA, and Perseus, which are used in many projects to improve semi-supervised methods specially addressed to classical and medieval literature. In addition, the growing number of platforms created to facilitate morphological analysis of new corpora, like Collatinus Lemlat, or GutenTag, is remarkable; these platforms offer a gateway to rapid extraction of a full range of linguistic attributes. 11

All these efforts are complemented by an increasing availability of programming packages including easy access to classical techniques like POS tagging, parsing, chunking and named entity resolution in combination with annotating tools based on treebanks-training such as TreeTagger and Lapos. They are contributing to the incorporation of NER in mid-range projects. Stanford CoreNLP, Freeling,, Natural Language Toolkit and Scikit-learn are among the most used set of tools and libraries. 12

Concerning the task of disambiguation (entity linking), projects are close to semantic Web and linked open data principles, based on interoperability and machine-readable data. Natural languages are ambiguous by definition, and retrieved entity mentions must be clearly identified. There are several propositions, especially from library domains, calling to use lists of authors and authoritative databases to link recovered entities with unique referents stocked in big databases like DBpedia [Lehmann et al. 2013], Freebase [Bollacker at al. 2008], or Yago [Hoffart et al. 2013] using URI's (Uniform Resource Identifier) [Nouvel et al. 2016] [Rizzo et al. 2011]. Some very interesting works based on small-size corpora have been taking this direction, especially studies by literary scholars [Frontini et al. 2016] [Elson et al. 2010]. However, they are limited to databases containing knowledge about relevant people or mapping modern texts. For instance, these works are tightly focused on geographical disambiguation, because places do not show a high number of possibilities. 13

To address both problems, automatic entity detection and entity disambiguation, it is necessary to have large, hand-made annotated corpora whose production is highly time-consuming work. For instance, computer acceleration is the most promising solution for dealing with the large and growing number of available digital historical documents. 14

Algorithm-based techniques can do intelligent tasks and provide a very valuable result. However, experience shows that high context dependency at different levels compels the combination of the automatic result with an academic methodology to define correct heuristics tasks and to incorporate the most adequate epistemological rules.

3. Corpora and Model

3.1. CBMA corpus description

The main corpus we use comes from a database of medieval Burgundian charters provided by the CBMA group.[1] The size of the complete database is about 29,000 documents stocked in cartularies and collections of originals, among which a sub-corpus of 5,300 items has been annotated with both person and location entities by the CBMA scientific team. The document set is mainly comprised of private charters produced between the 10th and 14th centuries in Cluniac abbeys, and a small part in Cistercian abbeys. The items come from almost 100 small places in Burgundy, and they are stored in ten different cartularies (mainly in four: Cartulary A and B from the abbey of Cluny, the cartulary of St Vincent of Maçon, the cartulary of the priory of Jully-les-Nonnains, and the cartulary from the Cistercian abbey of Vauluisant).[2] These cartularies were edited, during the 19th and 20th centuries, with different diplomatic and philological editorial standards. (Figure 1 shows how these acts spread over time.) 15

Typewritten texts digitized in modern editions are the primary source of available textual corpora, where elements such as capitalization, punctuation, and development of abbreviations have been added, therefore modernizing original sources for easier reading. Plain text has been stored in a database, and a team of historical and philological experts has manually annotated named entities. Due to lack of time and resources, personal and geographical entities have been tagged, but not juridical and institutional entities. 16

Our gold-standard corpus composed by the CBMA team is mainly distributed through five main types of acts: diplomas, charters, bulls, notices, and census lists forming a corpus of 5,300 items with 1.2 million words and almost 85,000 named entities. 17

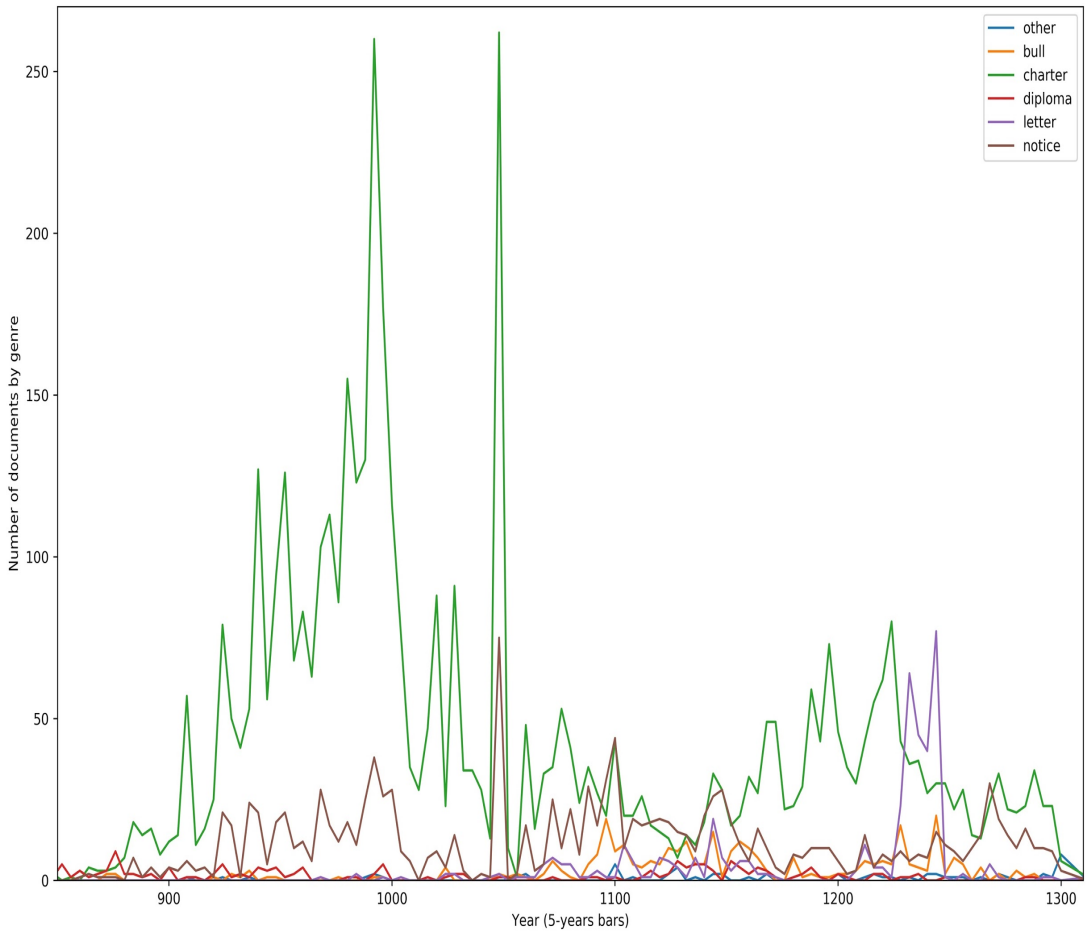

Figure 1. Chronological distribution and legal act classification in the CBMA annotated corpus

The importance of cartularies in medieval history is well known, for they gather transcriptions of original records and copies from various kinds of acts: royal deeds, privileges, judgments, private business, donations, etc. Such sources prove to be formal documents written by following formulaic patterns, which provide them with a more or less stereotyped structure. In these discourses, the names of people, places, ecclesiastical or other organizations, dates, titles, etc., are specific elements connected to the particular context of writing. All the documents are written in medieval variants of Latin, which uses special rules and vocabularies, and presents much more linguistic variation than classical Latin. 18

An example here of these documents may be enlightening. Dated between 994 and 1007 A. D,[3] this donation act to the Cluny abbey begins with a short protocol followed immediately by the dispositive indicating the object and purpose of the donation and the participants in the act: 19

> *In nomine Verbi incarnati. Cuncti noverint populi, quod ego Adalguis dono Domino Deo et sanctis apostolis ejus Petro et Paulo, ad loco Cluniaco, ubi domnus et reverentissimus Odilo abbas magis videtur preesse quam prodesse, pro redemptione omnium peccatorum meorum, ut Dominus dignetur auxiliare in extremi diem judicii.* 20

[4]

Coming up next is a detailed description and location of lands and properties that are the object of exchange in the donation act. The descriptions always involve nominal reference (metonymy) and they follow a hierarchy in land organization, in this case, *comitum, villa, terra*, inside which we can find the land properties classified in specialized parcels, here *campum, vinea, curtilus, pratum,* etc.: 21

Sunt ergo ipsas res in comitatu Matisconensi, in villa Tasiaco: hoc est in primis unum curtile cum domo et vinea, et campum; et habet fines de tres partes, de una terra Rodulfo, de alia Seguini, de tertia terra francorum, de quarta via publica. In alio loco alium campum; habet fines de quattuor partes, de una parte terra Rodulfo, de alia terra Fulchardo, de tertia via publica. Et in aliis duobus locis duabus peciolas de prato; habent fines de tres partes, de una parte terra Sancti Quintini, de alia Sancti Martini, de alia via publica.

Closing the dispositive is a *sanctio* containing penal clauses designed to reinforce and guarantee legal action:

De hanc donationem faciunt rectores de Cluniaco quicquid facere voluerint ab hodierno die. Si quis vero ullus homo donationem hanc contrariare voluerit aliquam litem, ira Dei incurrat super illum, et sit demersus in infernum vivus, si ad emendationem non venerit.

Finally, there is a two-fold eschatocol with the subscriptions of the donor and of witnesses, containing their names and an “S.” for *signum*(signature):

S. Adalguis femina, qui hanc cartam fieri jussit et firmare rogavit. S. Ingelelmi. S. Arlei. S. item Arlei. S. Bernardi. S. Ebrardi.

The writing style, the size and purpose of acts continuously vary, but they are follow a similar formulaic pattern. Except in long preambles – part of the protocol – where the writers of acts can express ideological motivations of actions by using quotation of authority texts, the acts do not include personal or narrative details. That gives the impression that we are facing a repetitive and modularly constructed source. This *dryness* is a necessity in documents that aim to become proofs of rights. The formulary pattern shapes the notarial written expression, and it is the named entity that specifies and individualizes the document. As we can notice in the above example, conventional structures also include hierarchical and referential elements and provide support for a dynamic use of named entities in each part of the charter. Indeed, the recovery of one named entity can involve the recovery of one piece of a complex regional landscape, the name of one tenant of a shared property, one node of a familiar or communal network, or one named participant in a long series of ecclesiastical transactions. Put together, these pieces can help to rebuild topographical indices and villas layouts, to propose social, juridical and economic frameworks established around land possession, or to describe a more detailed image of economic life in abbeys. From this, we can easily imagine the power of detection and recovery of named entities to quickly and extensively examine large databases of medieval documentation.

Conducting research using NER with historical corpora is not more complicated than working with Newswire or literature. Indeed, the different degrees of difficulty to complete a recognition model in any of these three areas comes from the varying degrees of observation necessary to detect main phenomena affecting the extraction of linguistic features. Historical literature about Cluniac cartularies is not lacking, and a bibliographical inspection can prevent some major problems affecting the collection of named entity occurrences. Four main highlights must be reported.

In the first place, because corpus goes from the late 9th century to the mid-13th century, we span the process of anthroponymic revolution in western Europe that begins in the early 10th century and extends until the 13th century. This is a process very well described by recent studies [Bourin 1996]. Roughly speaking, the single name denomination, a reduction of the Roman format used since early medieval times, was replaced around the late 10th century by a new tradition using the bi-name with two or more components (usually fathers’ names and locatives), and this new tradition was consolidated throughout the 11th and 12th centuries. From the mid-12th century on, this formulaic denomination adopted more complex modes using periphrasis, familiar references, professions, regional origin, even surnames, to form personal names with three to six components. The increasing irregularity of names represents one significant challenge to our model.

In the second place, we are often faced with overlapping and ambiguous series of entities which arise from three main documentary issues:

(1) the rise of the *donatio pro anima* formulary model around the beginning of the 10th century. This charter model

testifying to a property donation to the church takes place within the alms-giving doctrine [Magnani 2002]. Because cartularies gather together ecclesiastical property documents, a substantial part of our corpora follows this model. With the donation, the donor gives a property to a saint, who acts as an intermediary, and not directly to the church, and this property becomes part of lands under a saint's *dominium*, which is managed by a monastery or an abbey acting as a juridical person. In this panorama, charters contain extensive lists of named entities mixing personal, geographical and institutional functions under one name (ex. *rebus Sancti Vincentii Matiscensis, terra de Sancta Maria de Optimo Monte*, *in honore Sancti Petri constructam*, etc.). These nested entities, classified as place names rather than as names of persons, introduce a substantial bias factor to the model (See figure 2).

(2) a significant number of toponyms, starting from the 11th century, occurring as part of the bi-name personal denomination (acting as a locative complement). The overlapped entities usually have two or three components (ex. *Lambertus de Malliaco, Stephano de Cave Rupe*), but in the 13th century it is not rare to find entities with four or five components (ex. *Guillelmus de Sancti Stephano de Ponte*). This is a difficult technical problem, because disambiguation needs an advanced model able to classify one occurrence into two or more classes, and that is not always possible. If we take into account only the largest entity (in these cases, a personal name), we could lose thousands of associated micro-toponyms. 32

(3) the rapid extension of a common vocabulary in cartularies to describe lands, properties, and goods, which tended to fix spatial descriptions inside a stereotypical textual model. The textual environment uses and reuses concepts, *formulae*, and word associations arranged in a discourse containing parts usually well distinguished (formulary). This can be an essential factor of recognition, but it can also become a source of overfitting and overgeneralized training on co-occurrences. 33

In the third place, the production of the charters covers at least three periods of change linked to profound transformations in territorial reality, in the organization of writing, and in the legality of acts. Charters from the 10th century, which almost entirely concern ecclesiastical patrimony, are very close to the formulary; personal changes and adaptations are almost absent. They present long preambles and very precise and descriptive *dispositio*. In the early 11th century, even if the topics of charters do not change, they are partly replaced by *notices*, which are summaries of the charters drawn up more freely. Over the course of this century, numerous transformations in written vocabulary relate to profound changes in the spatial and legal structures affecting the redaction of charters [Bange 1984]. The broad nature of these changes in Western Europe, linked to feudalism, fuelled in 90's a historiographical debate concerning the *mutatio* of the year 1000 [Barthélemy 1997]. 34

In the fourth and last place, documents present a particular state of the language. Our four single-century corpora show linguistic features with a high degree of heterogeneity, because the documents were written during the course of a vernacularization process with progressive emancipation from the rules of classical Latin [Brunner 2009]. Much of the inconsistency and arbitrariness encountered during construction of an automatic model comes from this issue. The task is even more complicated if we consider that scarce NLP resources for Latin are adapted to literary texts based on classical Latin and devoted to the extraction of literary or philological features [Passarotti 2014]. In fact, the model proposed in this article tries to fill a major gap in this NLP research. 35

3.2. Corpus Modification and Extension

Modifications and adaptations of the CBMA corpus were motivated by questions related to the representativeness and robustness of our models. We aim to develop a model based on a regional corpus and to investigate whether it can be applied to another corpus created in different circumstances or belonging to different areas and traditions of writing. 36

There are three reasons why we created multiple sub-corpora in our study: 37

(i) First, *scriptoria* are localized in the Burgundy region, where Cluniac and Cistercian abbeys were founded, and where they developed their influence during the 10th, 11th and 12th centuries. 38

(ii) Secondly, the chronological distribution of the corpus is concentrated on the 10th and 11th centuries, which are the 39

period of the most substantial Cluniac production [Iogna-Prat et al. 2013]. In our corpus, the documentation before the 10th century is very scarce, and Latin is no longer the common language after ca. 1260.

(iii) Third, the charters have been obtained from ten different editions and twelve different cartularies concerning legal transactions from almost 100 small areas. 40

In other words, we are confronting a problem that is double in scope: on the one hand, we have a central institution (Cluny) that is the recipient of the documents and that applies certain corrections and alterations to them when making copies. Because the documents come from different areas with different scriptural traditions and preferences, the differences are largely authorized, and not the result of error. On the other hand, the over-representation of Cluniac and Cistercian charters can hide minor stylistic variations coming from smaller institutions. 41

In order to study the impact of all of these issues on the quality of our model, we performed different cross-validation and evaluation experiments; each one focused on a single, isolated aspect: corpus size, temporal variations, and regional variations. 42

3.2.2. Generating sub-corpora

Several sub-corpora were produced during the course of our experimentations by modifying the number of documents and the temporal distribution: 43

- The reduction of the size of the training dataset generally affects the accuracy of the model when used upon new data. We tested different sizes of training corpus to estimate the minimal amount of training data that is necessary to obtain good results. With our 5300-document corpus, we experimented with different sizes of training corpus: 4000, 2500, 1500, 1000 and 500 documents (we reserved a fifth of the documents in each set to use as a test corpus), and we applied the same training protocols in all cases. These experiments are important not only to the constitution of an extensible model, but also to the formation of a lighter model requiring lower computational resources.
- On the other hand, the purpose of temporal variations is to build a cross-validation environment for testing the robustness of the model on different chronological units, and then validating, or not, the application of the model to a wider temporal range. For this reason, we built century models, that is, learning models using only documents from the same century. The accuracy of this experiment may be questioned because, in many cases, the date of a document's creation is not certain, and the document is dated, then, using a time interval. This problem was corrected by building sub-corpora, considering a charter inside two corpora if its estimated date was near a transition period. For example, a charter dated around 980-1020 was considered to be in both the 10th- and 11th-century models. Thus, we obtained a biased result in contiguous model comparison due to the portion of shared charters, but, in compensation, we obtained sub-corpora that were more representative of each century, since changes of style or scriptural practices do not follow rigid chronological divisions. The distribution of all these corpora and sub-corpora can be observed in our Table 1.

3.2.3. Manual annotation of extra corpora

Nevertheless, both types of the previously mentioned experiments decrease the variety in the training corpus, and the features provided to the model can hide specific scriptural phenomena. To investigate this issue, an extra corpus of 400 items extracted from unlabelled data was tagged, with named entities covering the grey chronological areas of the 9th, 12th and 13th centuries. By adding this additional corpus to our sub-corpora, we expected to assure a chronological variety of smaller corpora and to avoid the loss of scriptural varieties. For that reason, documents were chosen in the decades that were the least represented in the original corpus. This supplementary corpus was also divided by centuries and each part was added as an auxiliary group to the century corpora. 44

Taking it even further, corpora and sub-corpora models were trained using charters from the same corpus. The robustness of these models had to be compared using non-Burgundian charters. To accomplish this task, four new 45

small-sized corpora were annotated by hand. These documents were extracted from Parisian, English, Lombard and Castilian charters from four regions with an intense charter production displaying different scriptural models. They are part of four of the most complete medieval charter corpora available online: the Ile-de-France cartulary (12th -13th centuries) published by the École des Chartes; the DEEDS corpus (10th-13th centuries); the CDML (*Codice diplomatico della Lombardie Medievale*, 11th-12th centuries) containing ecclesiastical and chancery charters; and the CODEA (CHARTA) corpus (10th-18th centuries) containing documents to map the rise of medieval Spanish.

The limited size of these sub-corpora (from 70 to 100 charters) made sub-corpora formation by century impossible in these cases, but we tried to balance them chronologically and to include not only cartulary documents but also some diplomas, bulls and administrative acts to increase heterogeneity. 46

Century/Corpora	ENG-LAND	CASTILE	ILEFRANCE	LOMBARDY	Original Corpus	Modified Corpus	400_Set
10th	10	7	9	8	2292	3230	12
11th	24	13	23	17	1510	2050	27
12th	24	16	54	23	816	860	182
13th	12	14	63	2	638	730	149
Nº Tokens	11110	15616	41608	12441	1096095	1096095	104330
Nº Entities	1326	1841	3594	1222	84752	84752	8263

Table 1. Table 1: Nº of charters for each century and global features in gold-standard corpus and external corpora

3.3. CRF modeling

3.3.1. Normalization and segmentation

This section describes the pre-processing steps applied to all our corpora. First, we did a manual validation of doubtful items in the training data, mostly for detecting nested, overlapping entities and ill-formed annotations. For example, the description of territorial boundaries with abuse of possessive genitive (or nominative) forms leads to overlapping entities, but also to entities with unclear boundaries, and is due to a much freer order in the phrase and use of declensions than in classical Latin (see examples in Table 5). Overlapping entities are usually connected to names serving different functions, frequently saints' names. They can refer to a personal name or the name of an abbey, a piece of land, a feudal territory, or even a festivity date. On the other hand, the texts from philological and diplomatic editions can contain paratextual data such as special characters, glosses, titles, etc., that can introduce noise into the model. A process of normalization in textual spelling is completed through scripts. This normalization decreases error rate and makes the transformation of the original format corpus easier. 47

We then split the main corpus into three sets necessary for the creation of the model, i.e., training set, development set, and test set. We randomly selected a training corpus of 4,000 documents (>one million words), a test corpus of 1,000 documents, a development corpus of 300 documents forming the development set, used to avoid overfitting, i.e. a model that is too close to the test set and thus not robust to new data. To ensure the chronological presence of documents in each sub-corpus, we also tested semi-random distribution by creating clusters of documents every 25 years, but this approach did not show significant differences in the results. That is why we preferred the random method in order to maintain the original distribution of data. The temporal distribution of the three datasets turns out to be very similar to the distribution of the entire corpus. 48

3.3.2. Model

We see our NER problem as a problem of traditional sequence classification, and we use the well-known BIO format to represent the data. In this format, each token (in our case, each word) is assigned to a BIO class: B-entity, I-entity, and O-entity, respectively, representing the beginning (B), continuation (I as "inside"), or absence (O as "outside") of named 49

entities (See Table 2). We then apply Conditional Random Fields [Lafferty et al. 2001], one of the most popular methods for this category of problems. To apply this supervised machine-learning approach, each word in a sentence must be regarded as a token. The entire corpus was converted into a tabular format, providing lexical, syntactical, and morphological information at a token-level. Each word was represented by the following:

- TOKEN (original word)
- POS (part-of-speech category)
- LEMMA
- CASE (whether the first letter is in upper or lower case)
- SUFFIX (last three letters of the surface form)

The first three features (TOKEN, POS and LEMMA) were obtained from a version of the tool TreeTagger adapted to Medio Latin, created by the OMNIA group in 2013. These three features help the model to consider the grammatical and morphological information from the text. The next two columns (CASE and SUFFIX) help the classifier to exploit capitalization and the inflected nature of Latin, in which the end of the word determines its grammatical function. 50

We considered the problem as a two-step classification: the first step extracts the personal names, while the second step extracts location names. A single classifier extracting personal and location names jointly could not be implemented, as the corpus contains a lot of overlapping entities. That is why the last columns of Figure 2 list the classes in BIO format. 51

		LOC			ORG		
PERS					PERS		LOC
Hugonis	de	Breza	donat	monasterio	Sancto	Petro	Cluniacensis
NAM	PRE	NAM	VBE	SUB	QLF	NAM	NAM

Figure 2. Example of two and three nested levels of overlapped entities

The CRF method operates by forming a discriminative model and finding the best state assignment from a set of training corpora containing tagged features. That is, given a series of labelled observations it constructs an interpretation and from there it determines probabilistically the most approximate label for a new unseen sequence [Wallach 2004]. As explained above, the corpus has been divided into three different sections: training set, validation (development) set, and test set. After each iteration on the training set, the learned model is applied to the validation set, and the process is stopped when the validation result does not improve anymore (“early stopping”). The test set was reserved to evaluate the performance of the model with data that did not participate in the learning phase, thus avoiding any bias. 52

The features used by the CRF model are specified by a list of patterns determining the observation rules for each element (word) of the sequence (document). A token feature can be one of the seven columns on the token row, the rows before or after, or a combination of these. This allows using the immediate context of each word. The pattern that we have written combines unigrams and bigrams in an extended sequence of grams (i.e., regarding words one by one or two at a time), combining two forward and two backward token-line positions. An efficient feature selection and induction by the L-BFGS algorithm was provided by Wapiti [Lavergne et al. 2010] in a sequence-labelling toolkit developed at LIMSI-CNRS. 53

TOKEN	POS	LEMMA	CASE	SUFFIX	ENTITY	ENTITY
Quod	CON	Quod	UPPER	Uod	O	O
ego *[2,0]*	PRO	Ego	LOWER	Ego	O	O
Hugo *[1,0]*	NAM	-	UPPER	ugo	B-PERS	O
de *[0,0]*	PRE *[0,1]*	de *[0,2]*	LOWER *[0,3]*	de *[0,4]*	I-PERS *[0,5]*	O *[0,6]*
Berziaco [-1,0]	NAM	-	UPPER	aco	I-PERS	B-LOC
perpendens [-2,0]	VBE	Perpendeo	LOWER	ens	O	O
,	PON	,	LOWER			

Table 2. Table 2: Training sample for the sequence “Quod ego Hugo de Berziaco perpendens.” The gray zone indicates one single observation (i.e., the word de) combining features from all columns in a window of five tokens (from two tokens before to two tokens after the observed token).

4. Evaluation

4.1. Model experimentations

All models were evaluated with traditional precision, by recall and F1-measure with two different configurations, as implemented in the tool BratEval: “exact match” counts a true positive if an extracted named entity has the correct type (person or location) and if the boundaries of the extracted entity match perfectly with the gold standard. On the other hand, “partial match” counts a true positive even if the extracted entity shares only a partial overlap with the gold standard. 54

4.1.1. General model

We first ran our model trained with the entire training and development CBMA set, regardless of its temporal and geographical characteristics. This model achieves an exact match F1-measure of 0.95 and a partial match of 0.96 on test data for people’s names; it achieves an exact match of 0.91 and a partial match of 0.92 for location names. All experiments show a similar result in precision and recall parameters. The difference between exact and partial results is not more than two points, confirming that boundary detection is not very problematic, although it is often a hard task in NER. 55

Personal name	PR	RC	F1	TP	FP	FN
B-PERS	0.95	0.96	0.96			
I-PERS	0.88	0.92	0.90			
Partial match	0.95	0.97	0.96	12965	615	291
Exact match	0.93	0.96	0.95	12729	851	529
Location name						
B-LOC	0.91	0.93	0.92			
I-LOC	0.81	0.80	0.80			
Partial match	0.92	0.92	0.92	7171	590	550
Exact match	0.90	0.91	0.91	7035	726	681

Table 3. Table 3: Best current ratio recognition expressed in BIO-tag ratio of recognition and number of occurrences according to Brateval tool which provides a pairwise comparison of annotation sets in Brat format. The tool displays two results: Exact match (EM)when entities agree in type and extension, and Partial match (PM) when entities agree in type but not in extension. True positive (TP), False Positive (FP), False negative (FN), Recall (Rc), Precision (Pr) and F1-measure

4.1.2. Century models

Tables 4 (personal names) and 5 (location names) show the result of our cross-validation experiments on century models. For each one of the four considered centuries (from the 10th to the 13th centuries), we applied a model trained on one century to another century set. Cross-results did not show a considerable heterogeneity, and the results are very close to those obtained with big-sized models. The main differences are detected while evaluating datasets and models formed on the basis of charters from the 10th and 13th centuries. Here there are three highlights: first, the 10th-century model constructed from the largest dataset gives us the worst results when applied to the other charters, especially to the 12th- and 13th-century datasets; second, the application of the 10th-, 11th-, and 12th-century models to the 13th-century dataset offers the lowest compared performance; third, the worst individual performance is obtained by crossing the 10th-century model with the dataset of the 13th-century charters. 56

All this suggests two major observations: (i) there are no significant gaps in the variety of features of corpora modelled on four centuries of charters, thus maintaining a regular detection of naming phenomena that produces results similar to those of single-century models; (ii) in consequence, detection problems are concentrated at a feature-level in irregular denominations and in severe changes in the name's composition (or in its textual disposition) that have taken place in the 13th but not in the 10th century or vice versa. 57

4.1.3. European charters

Tables 6 and 7 (corresponding to names of persons and locations) show the results obtained by the application of our different models to the “foreign” corpora described in section 3.2. We apply to these regional corpora the models trained with the general CBMA datasets (two models with different sizes) and the four, single-century sub-corpora trained from the same dataset. The six results obtained show great proximity among them. General models provide better coverage, normally around 3%-5% superior to the century models, and again, we can see that the model trained on 1,500 charters shows a small decrease in performance (1-2 points) compared to the full model. The use of a smaller set of annotated documents could be acceptable. 58

The performance of general model is quite appropriate: PERS partial-match recognition is between 94% and 93%, and exact-match recognition shows results between 87% and 81% for the four regional corpora. In the case of LOC entities, the partial-match recognition reaches results between 85% and 82%, and exact-match recognition is between 81% and 73%. Analysing only the numerical results from single-century models, we noticed many similarities to the results obtained in previous century model evaluations: (i) the best performance models are the central models from the 11th and 12th centuries; (ii) boundary models (from the 10th and 13th centuries) are less adequate for recognition, especially of compound entities; and (iii) the model from the 10th century, trained with the largest number of documents, offers the lowest efficiency on foreign corpora recognition. 59

We also noticed a correlation between single-century models and regional corpora. For example, the Parisian corpus, composed mostly of later medieval charters, offered slightly higher numbers when evaluated with the 12th- and 13th-century models; the Lombard corpus, concentrated in the 11th and 12th centuries without any 13th-century charters, showed lower performance when evaluated by 13th century model, while the English and Hispanic corpora, with the most balanced chronology (the Hispanic corpus is not really an organic corpus), presented greater distance between partial and exact matches when evaluated with all four of the century models. 60

To summarize, the similar results obtained from the six models help to reinforce the hypothesis about the results of century model evaluations: (i) there are long series of features in documents that resist changes in the chronology, size and origin of the charters, features that are better treated by an iterative model based on the formulary and on established scriptural traditions, assuring thereby an initially acceptable ratio of recognition; (ii) there are more discrete sets of features, added to previously mentioned ones, determined by local specificities, document changes and highly specific phenomena, which become the primary source of errors. In addition to these, the set of particularities from the four external corpora were treated without a considerable decrease in the general performance, suggesting that there are global similarities in the composition of the name that are much stronger than differences suggested by each regional tradition. 61

Model/ Test		ANGLO		CASTILE		ILE_FR		LOMB		10th		11th		12th		13th	
		Pers	Loc	Pers	Loc	Pers	Loc	Pers	Loc	Pers	Loc	Pers	Loc	Pers	Loc	Pers	Loc
5000	PM	0,93	0,82	0,93	0,85	0,94	0,85	0,93	0,84								
	EM	0,83	0,76	0,81	0,73	0,87	0,79	0,85	0,81								
1500	PM	0,89	0,82	0,92	0,85	0,94	0,85	0,93	0,83								
	EM	0,83	0,77	0,80	0,73	0,88	0,79	0,88	0,79								
10th	PM	0,93	0,66	0,92	0,75	0,89	0,75	0,93	0,72			0,97	0,89	0,94	0,86	0,84	0,83
	EM	0,86	0,60	0,75	0,63	0,82	0,69	0,87	0,65			0,94	0,87	0,85	0,81	0,75	0,76
11th	PM	0,94	0,80	0,92	0,83	0,91	0,83	0,93	0,83	0,98	0,93			0,96	0,93	0,86	0,88
	EM	0,88	0,74	0,77	0,70	0,85	0,76	0,88	0,78	0,97	0,91			0,90	0,89	0,79	0,82
12th	PM	0,91	0,79	0,88	0,84	0,92	0,85	0,89	0,75	0,96	0,88	0,96	0,90			0,88	0,89
	EM	0,83	0,75	0,76	0,73	0,86	0,79	0,84	0,70	0,95	0,85	0,95	0,88			0,81	0,86
13th	PM	0,84	0,71	0,81	0,77	0,91	0,84	0,85	0,60	0,94	0,87	0,93	0,84	0,92	0,89		
	EM	0,71	0,78	0,69	0,65	0,83	0,78	0,78	0,56	0,93	0,85	0,90	0,82	0,84	0,86		

Figure 3. Table 4: Accuracy scores obtained on cross evaluation for names of persons and places using century and general models on European and century set test.

4.2. Match and error analysis

High performance on a partial match means that the model detects the presence of named entities no matter what size they are. In the case of single-named entities, there is only one single target element, and we can assume that the results in the detection are almost the same as the classification of the entity. In the case of complex entities (two or more components), complete classification is less efficient than detection by around 5-10 points (the difference between partial and exact matching), but there is still high performance at the task of recognition. 62

These results can become more meaningful if we remark three significant situations detected concerning named entities in our corpus and in evaluated corpora: 63

(1) As we mentioned above (section 3.1), the most widely diffused type of complex name is formed by the association of two or three single entities. In fact, around 86% of the compound personal names in the CBMA corpus have between two and three parts. This denomination format is generally composed of *name* + *de* + *toponym* (ex. *Bertrannus de Verziaco*). But in other cases, this double name format can be produced though declension, syntagmatic or periphrastic denominations, or even adding a second component using a genitive or nominative. There is a juxtaposition of two single names in the *name* + *name* (or *noun*) or *name* + *nexus* + *name* (or *noun*) formats (ex. *Rudericus Didaci, Bellonus Mangacii, Gariardus de loco Antimiano*). 64

(2) From the 11th century on, more than half of the geographical entities detected form part of a personal name. The *name* + *de* + *toponym* format enjoyed great success until the 13th century. In this format, we generally consider the last name as a location indicator (usually a micro-toponym or hagio-toponym), which means that well-performed boundary detection for a double name involves full geographical entity detection (that is a B-PERS + *de* + B-LOC pattern). 65

(3) Complex and single personal and geographic entities are closely related to an extensive and precise vocabulary of co-occurring words. This is a necessity in legal texts about properties where donors must be clearly identified, properties must be well-described, and signatures and signs of validations must be included. Lands and people are identified by using terms and titles of presentation expressed in a common vocabulary that responds to a spatial and social reality, but also to the uses signified by the formulary. 66

According to the evaluation results, the model can detect the apparition of entities with up to 95% efficiency and classify them with an accuracy of up to 90% in the case of single entities as well as in the case of the most extended forms of composed entities: the double locative name and the double name using the *de* particle, the genitive, or any other single nexus. 67

In this respect, we can confirm a model able, on the one hand, to classify with high performance both the single names used almost exclusively until the 10th century and the most widely diffused name format in medieval times since the 11th century: the double name in all its variations. We can also confirm a model able, on the other hand, to recognize 68

the non-complex entities associated with a large, social, and professional vocabulary. (See Table 4.)

69 Furthermore, as long as the double name maintains a two- or three-part compound format, the model has a high level of efficiency (90% or more) due to limited variations and not very complex n-gram observations. Providing information related to lemma and termination (or declension), capital letters and grammar features is usually enough for the model to accomplish named entity recognition on compound names.

70 Taking that into account, the most frequent errors in recognition must necessarily be linked to long, complex denominations and specific phenomena such as grammatical forms in the Latin phrases or to changes in underlying textual structures. In Table 5, we can see displayed some examples of frequent errors made by the system. For example, in Lombard texts the prevalence of periphrastic denominations produces a lower ratio for the recovery of inside parts of named entities because of the high frequency of non-entity words between names (see examples 1 and 2). In 11th- and 12th-century charters, the increasing use of the genitive in naming entities produces a large group of entities without particles. In addition, the coexistence in medieval Latin of the inflectional rules of classical Latin, the expansion in the use of prepositions, and the gradual extinction of Latin cases can all lead the model to some confusions (see example 4).

71 It is especially hard to handle the apparition and rapid expansion of institutional names under saints' avocations in diplomatic donation models in the 10th century due to overlapping between personal, institutional, and place names (examples 6 and 8). The model does not work well with overlapped entities because most machine learning classifiers are not designed to attribute more than one class to each instance. In that sense, the confusion between place names and personal names must be solved by designating one class for each entity.

72 The overgeneralization of very common particles (such as *de* in compound entities), as well as of location trigger words (such as *terra, serum, pars, domus, manus, apud*) and also of personal co-occurrent words (such as *episcopus, beatus, dominus, sacerdos, miles*, etc.) can lead to false positives when the model finds an entity different than that expected, as illustrated by examples 5 and 7. On the other hand, because of the more flexible nature of the phrase order in Latin, the system can classify as inside entity (I-PERS, I-LOC) those nouns that are single-named entities or non-name entities.

73 Finally, in the course of all these experiments, we detected an important number of errors related to the original format of the documents. These texts are not raw data, and they contain many added features characteristic of a philological, palaeographical, or diplomatic edition, such as special characters, textual gaps, abbreviations, headlines, and orthographic redundancy (example 9). Not all this information is relevant to the model, and this can produce an important amount of noisy training data. Automatic and manual cleaning can resolve a significant part of this problem.

1) Lombardy, 1198				*2)Lombardy, 1145*				*3) England, 1082*			
Andreas	O	B-PERS	B-PERS	Ottonis	O	B-PERS	B-PERS	Gualtero	O	B-PERS	B-PERS
Budellus	O	I-PERS	I-PERS	qui	O	O	I-PERS	de	O	I-PERS	I-PERS
de	O	O	I-PERS	dicitur	O	O	I-PERS	Monte	B-LOC	I-PERS	I-PERS
loco	O	O	I-PERS	de	O	O	I-PERS	Sancte	I-LOC	I-PERS	O
Cossonno	B-LOC	O	I-PERS	Suso	B-LOC	O	I-PERS	Trinitatis	I-LOC	I-PERS	O
4) England, 1067				*5) Ile de France, 1237*				*6) Ile de France, 1116*			
Willelmus	O	B-PERS	B-PERS	domino	O	O	O	supradicta	O	O	O
Dei	O	O	I-PERS	Petro	O	B-PERS	B-PERS	beati	B-LOC	O	O
gratia	O	O	I-PERS	milite	O	O	I-PERS	Martini	I-LOC	B-PERS	O
rex	O	O	I-PERS	de	O	O	I-PERS	de	I-LOC	I-PERS	O
Anglorum	B-LOC	O	I-PERS	Sarrangi	B-LOC	O	I-PERS	Campis	I-LOC	B-LOC	B-LOC
Willelmi	O	B-PERS	I-PERS								
7) Castile, 993				*8) Castile, 1246*				*9) Ile de France, 1086*			
que	O	O	O	pro	O	O	O	S.	O	O	O
uocitant	O	O	O	monasterio	O	O	O	Hi	O	B-PERS	O
Arroyo	O	B-LOC	B-PERS	Sancti	O	B-LOC	B-PERS	[	O	I-PERS	O
de	O	I-LOC	I-PERS	Emiliani	O	I-LOC	I-PERS	uonisa	O	I-PERS	O
Sancti	B-LOC	I-LOC	I-PERS	de	O	I-LOC	I-PERS	]	O	O	O
Fructuosi	I-LOC	I-LOC	I-PERS	Pledanos	B-LOC	I-LOC	I-PERS	S.	O	O	O
								Roberti	O	B-PERS	B-PERS

Figure 4. Table 5: Examples of frequent errors made by the system on European charters. The first column of each example shows the text excerpt; the second and third column shows the gold standard for location and person names, and the last column shows the system automatic classification.

5. Discussion

In this work, we use different types of information to build a model for automatic named entity recognition: grammatical categories, lemma, case, entity categorization and n-gram observations. A model based on a probabilistic CRF-approach that extracts correlations dependencies between this information and observations leads us to high performance in the process of categorizing entities. High percentages in the results, after having tested the model on corpora of different geographical origins and different periods, demonstrate that the linguistic and statistical algorithmic approach has proven to be robust. In fact, the results are quite conclusive, but the evaluation process involves an important level of heterogeneity needing an explanation that exceeds the most technical level and requires the use of humanistic knowledge. We need to answer why a model created from a regional corpus can obtain a high recognition rate on medieval documents from other regions, chronologies and even traditions. 74

Thus, references to the historical background, social usages and scriptural variations can provide a more meaningful explanatory context, enriching the evidence obtained by machine-learning methods. Reviewing the literature about historical names is a good first step. Anthroponymic studies suggest a double movement in Western Europe after the 10th century: first, a reduction of name stock and consequently a reduction of the variety of common names; and second, an extension of double names in variable forms: *nomen paternum* (name + name of father), locatives (name + place-name) and nicknames (name + profession, qualities or titles). In fact, the nature of the second element in bi-names distinguishes several naming traditions in Europe. In France, the meridional tradition incorporated the *nomen paternum* quickly, that is, the hereditary second name coming from the father, but central and northern regions (like Burgundy and Île-de-France) developed a preference for locatives and periphrastic forms [Bourin 1996]. 75

Single-name forms disappeared in the 12th century, and traditional and local names were replaced by the universal Christian or princely names. At the same time, in the northern Hispanic regions and Castile, the tradition promoted an incorporation of *nomen paternum*, formed with the declension in the original name of the father (v.g. *Gundisaluo Roderici, Rodericus Ferrandi*) and, later, a composition of three names in association with locatives and nicknames (v.g *Ferrando Sanchiz de Fenosa*) [Sopena 1996]. Norman traditions heavily influenced the central and southern regions of England, which adopted double locative or triple names and *nomen paternum* formed by the second name of the father (v.g *Guidonis Nerioli de Buxiaco, Aimo de Monte Pauonis*) [Billy 1995]. The tradition operating in the Lombardy region presents the same phenomena of duplication in anthroponymic elements during the 11th century, but prefers a system that promotes a more broadly referential second name (v.g. *Otonnis qui dicitur de Suso, Grioulum filium Lafranci Caipeni*) which takes the form of a periphrastic composition promoting surnames, geographical origin, and ancestor's filiations [Corrarati 1994]. 76

Concerning geographical names, we must make a distinction between landscape entities and nominal entities. The first are used to describe or to locate a property, whereas the second appear as a locative complement of a compound name. Indeed, a large proportion of geographical entities is linked to the rise of double names between the 11th and 13th centuries, when the locative complement became a form of familial or personal social distinction. The expansion and precise definition in Europe of the *name + locative* combination through use of the *de* particle (genitive or nominative without preposition) produced a very strong increase in binomial name entities combining personal and geographical components. This is a process that quickly produces great variety by proposing the origin, residence, feudal, or ethnic place as locative complement (v.g. *Iohannes Allemanus*) completing the information about an individual person. In fact, all these locatives are generally real places, but they are usually defined as hagio-toponyms, old toponyms, and micro-toponyms hardly connected with real spaces on a map. 77

Nevertheless, most of the geographical entities are not associated with a personal name but correspond to the first above-mentioned type: landscape and territorial names. They are the names of a *terra*, a *villa*, a *pagus*, an *ager*, a river, a building, a church, a monastery, a so-called place, etc. Names and master words or co-occurrences are closely related in charters (see Table 5) because cartularies produce, by classifying records geographically and transcribing them in a codex, a spatial knowledge of ecclesiastical estates which uses many of the same global geographical entities [Chastang 2007]. The first effect of this is the constant reference in formulary charters to a well-delineated territorial space, a cadastral order that today is almost impossible to reconstruct due to numerous information gaps [Bange 1984]. 78

Personal entities	**Geographical entities**
Professions and social activities: *camerarius, cantor, magister, miles, monachus, notarius, sacerdos*	Landform spaces: *boscus, fluvius, locus, mons, nemus, pratus, rivus, silva* 79
Secular and religious titles: *abbas, beatus, comes, domnus, dominus, dux, episcopus, papa, presbyter, princeps, rex*	Seigniorial and ecclesiastical division of sedentary lands: *ager, conventus, curtilus, domus, feudus, grangia, mansus, pagus, vicus*
Dignities and Nicknames: *benedictus, brunus, grossus, humilis, largus, normandus, paganus, servus, venerabilis*	Legal and jurisdictional division: *area, castrum, civitas, diocesis, ecclesia, provincia, sedes, terra, villa*
Periphrastic nexus: *appelatus, cognomen, dictus, nomen, vocatus*	Landmarks and micro-spaces: *altar, atrium, capella, capitulum, castellum, cenobium, domus, ecclesia, hospital, monasterium*
Words with nominal value: *alius, ego, filius, frater, idem, nepos, signum (S.), uxor*	Prepositions, adverbs or global locative value terms: *ad, apud, fines, inter, manus, meridies, pars, pro, supra, vocabulum*

Table 4. Table 5: Co-occurrences or presenting words (appositional, periphrastic and attributive nouns) of personal and geographical named entities. This table corresponds in the broadest sense to a jurisdictional and social vocabulary of personal and geographic categorization.

According to the evaluation of our general model, the results obtained in B-LOC recognition (geographical Begin-entity) are just five points below (90% compared to 95%) the best performance obtained in B-PERS (personal Begin-entity). That means that performance on detection of geographical entities and the classification of single entities is almost as excellent as obtained on personal names. Geographical entities usually appear in less complex forms than the entities of the name, which raises the percentage of this first result. But performance decreases almost ten points if inside entities (I-PERS and I-LOC) are compared (90% and 80%), which indicates that elements involved in composed geographical entities are harder to model. This proportionality is more or less maintained in evaluating foreign corpora (see Tables 6 and 7). All results are, on average, acceptable, but, once more, the problems are concentrated on inside-entities, as is shown when comparing partial and exhaustive Brateval results. 80

The most common format for complex entities is the double name (bi-name), but as we have seen, our model can handle well the recognition of this prolific format and its variants, which suggests that we need to examine more 81

complex forms of composed entities. Thus, regarding the apparition of complex geographical entities, we notice different situations depending on regional traditions. Concerning personal names, Castile and Southern England are regions where the triple name expansion takes place one century before it does in Central France. Both regions form bi-names with two nominal components, and the incorporation of locatives entails triple or quadruple names (v.g. *Guilielmi de Sancto Satiro*, *Martinus Garciez de Stallaia*). In Lombardy, the formation of complex names never takes root, but the addition of periphrastic composition (through *dicitur, nominatur, filius, de loco*, etc.) creates some long entities impossible to dissolve. Almost the same thing happens in central French regions where representative quantities in triple names are not present before the 13th century. But the use of locatives and periphrastic forms since the 11th century contributes to the apparition of complex entities of up to five elements (v.g. *Adalbertus de vico Camonaco filii Iohanni*).

Nonetheless, the most complex entity format is related to overlapped and nested entities (see Section 4.2). Personal entities, acting as possessive genitive or nominative, can be found fulfilling the function of a geographical entity in the land and boundary land descriptions in donations, purchases, inheritances, or legal disputes (v.g *a mane terra Bertrannus; sub domi ipsius Ansaldi*). Furthermore, since the 10th century, saints and ecclesiastical institutions were currently present in the *formulae* of the *donationes pro anima*, where they performed the function of intermediate receptors and defenders of properties. Their presence in charters contributes to creating complex entities with four or more elements taking, associating, or superimposing a saint's name, an institutional name, and a geographical name. 82

In part, this complexity is, even more, a result of medieval variants of the Latin language. The gradual expansion of use of the genitive since the 11th century created large groups of entities without intermediate particles; at the same time, there was an increase in prepositions and the extinction of many Latin case endings, all of which lead the model, in many instances, to recognize the words of non-entities as named entities. Latin phrase order is irregular, and exceptions in medieval variants are almost infinite; consequently, training taking into account grammatical rules, co-occurrences, and context can generate many false positives even with an approach that is not rule-based. 83

The difficulty in recognizing entities involved in these complex phenomena lies not so much in their quantity as in their extensive consequences. The percentage of complex entities (more than three elements) does not exceed 11% of the total in our corpora, but the statistical impact on results due to bad recognition of such entities is more elevated when they are composed of more than four elements: B-PERS + (3) I-PERS (that means three inside-entity errors for each non-recovered entity). Moreover, the impact of these phenomena is increased if we consider that the original mark-up of our database, in which non-physical personal names and juridical actors are classified as place-names, can elevate the number of false positives in automatic recognition of complex place-names. 84

The influence of these circumstances is more evident in the different tests of temporal variability that we have performed. On one hand, the results of century models applied to European charters are very stable, which confirms, first, the presence of important homogeneous structures in name compositions and formulaic patterns and, second, the presence of a larger stock of name compositions than we would have expected in the Burgundy region in each century, which provides more variety in the training process. On the other hand, in all cases, the model developed with charters from the 10th century, when the double name was not yet established, is less adequate for recognition in other centuries; at the same time, the models from the 10th, 11th, and 12th centuries are less efficient when used on 13th century charters, because larger and more complex compositions increased in this century. 85

Finally, the excellent statistical results in spite of temporal variability also show that there is a high degree of iteration in late medieval formulaic discourse. This characteristic helped us to form a valid model starting from an incomplete and not very structured corpus. We obtained an acceptable ratio of recognition when we tested the model on charters representing different legal acts: donations, transactions, royal orders, deeds and notarial actions from four European regions. That leads us to assume a vast radius of application for our model on Western European legal acts of various types. In addition, single-century models do not show substantial improvements compared to the results of general models when applied to documents from each century, which confirms that our general model can provide the best ratios of recognition along a temporal line of at least four centuries, from the 10th to the 13th century. 86

6. Conclusion

87 We present a named entities recognition model applied to medieval Latin charters, and we implement an adequate evaluation environment. The model produced from long data sets of Burgundian medieval charters is able to accomplish a high ratio of recognition of the personal and geographical names of medieval Latin documentary sources, especially from the 10th to the 13th century. The evaluation is focused on demonstration of the robustness of the model in diverse situations. We performed different cross-validation experiments modifying size, chronology, typology, and regional origin in several sub-corpora, obtaining a very acceptable ratio of recognition in each case, which confirms that our model has a wide range of applications with respect to medieval documentary sources. By crossing results, we can confirm that the model supports an accurate semi-automatization of named entities recognition and, in consequence, that it can provide a primordial level of structuration of data, saving many human efforts.

88 Validation and cross-experiments also show us that questions about representativeness in the learning model do not generate insurmountable issues. Our discussion tries to confirm that the most significant problems in recognition are an expression of the intense dependency of data on complex historical, social and scriptural conditions, and not only on linguistic or statistical concerns. Several historical and philological conclusions helped us to refine and normalize algorithmic approaches and to understand numerical results provided by the machine. This also led to corrections of the original dataset and to complement it by the integration of the new annotated datasets that we have produced during this project.

89 The use of this NER model can be fruitful in different fields of research in the Humanities, including, in particular, recovery information systems, automatic indexing and distant reading. Since named entities are not subject to a lot of variation, they can be used as keywords and meaningful terms improving internal workflows of search engines. A NER model which can automatically spot these keywords within a large dataset, becomes a useful tool to speed up, classify and refine requests on ancient documents that are available in plain text editions. The recovery of entities, acting as nodes of action, can serve as a first step in the production of historical GIS-maps, and the reconstruction of event timelines or social networks. These widely used forms of data visualization can benefit from automatic tagging while automatic tagging can also help to solve questions concerning the production or the dating of ancient texts by providing data that scholars have traditionally used in their textual or manuscripts studies – mention of persons, places, words or formulas. Since it enriches datasets with specific linguistic and semantic features, our NER model can even be integrated into a pipeline of tasks, such as topic classification of ancient texts, automatic handwriting recognition or word sense disambiguation.

* This project is supported by the "IDI 2016" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02

90

Notes

[1] The entire corpus is freely available in multiple formats on the website of the CBMA project: http://www.cbma-project.eu/bdds2/2014-07-10-14-27-56.html

[2] Bernard, Auguste & A.Bruel. Recueil des chartes de l'abbaye de Cluny. Imprimerie nationale, 1876; Petit, Ernest. Cartulaire du prieuré de Jully-les-Nonnains. Imprimerie de Georges Rouillé, 1881; Ragut, Camille; Chavot, Th. Cartulaire de Saint-Vincent de Mâcon: connu sous le nom de Livre enchaîné. Impr. d'É. Protat, 1864.

[3] BERNARD A., BRUEL A., *op.cit*, tome 3 : 987-1027, p. 245, n° 2039 (Collection de documents inédits sur l'histoire de France. Première série, Histoire politique).

[4] « In the name of the Incarnate Word. Let all the people know that I, Adalgis, donate to the Lord God and to his holy apostles, Peter and Paul, in the place of Cluny, where the master and most venerable abbot Odilon seems to rule more than to benefit, for the redemption of all of my sins, and for the Lord to assist me in the last day of judgment.There are, then, these things in the county of Macon, in the villa of Tasiaco: that is, in the first place, one *curtilus* (small estate) with a house and vine and field; and it has boundaries from three sides, from one Rodulfus' land, from another Seguinus' land, from the third Franks' land and from the forth the public road.Another field in other place; it has boundaries from four sides, from a side Rodulfus' land, from another Fulchardus' land and last one the public road. And two small pieces of meadow land; they have boundaries from three sides, from one side Sanctus Quintinus' land, from another Sanctus Martinus' land and from the other the

public road. The rectors of Cluny can do whatever they want, from today on, with this donation. However, if any person would want any legal dispute to stand against this donation, let God's wrath fall on him, and let him be submerged alive in hell if he doesn't return to amendment.Sign(Signum) Adalgis', woman, who ordered this act to be made and asked to validate it. S(ignum) Ingelelmi, S. Arlei, S. another Arlei. S. Bernardi. S. Ebrardi. »

Works Cited

Abacha et al. 2011 Abacha, A. B, Zweigenbaum, Pierre. (2011). "Medical entity recognition: A comparison of semantic and statistical methods." In *Proceedings of BioNLP* 2011 Workshop. Association for Computational Linguistics, p. 56-64.

Bange 1984 Bange, F. (1984). "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IXe-XIe siècles)." Annales. Economies, sociétés, civilisations, 39(03), pp.529-569.

Barthélemy 1997 Barthélemy, D. (1997). " La mutation de l'an mil, a-t-elle eu lieu ?" *En Annales. Histoire, Sciences Sociales.* Cambridge University Press, p. 767-777.

Billy 1995 Billy, P. (1995). "Nommer en Basse-Normandie aux XIe-XVe siècles." *Cahier des Annales de Normandie*, 26(1), pp.223-232.

Bollacker at al. 2008 Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. (2008). "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge." *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, ACM, pp. 1247-1250.

Bourin 1996 Bourin, Monique. (1996). "France du Midi et France du Nord : deux systèmes anthroponymiques? L'anthroponymie document de l'histoire sociale des mondes méditerranéens médiévaux." *Publications de l'École française de Rome*, 226(1), pp.179-202.

Brooke et al. 2016 Brooke, J., Hammond, A., Baldwin, T. (2016). "Bootstrapped text-level named entity recognition for literature." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pp. 344-350.

Brunner 2009 Brunner, T. (2009). "Le passage aux langues vernaculaires dans les actes de la pratique en Occident." *Le Moyen Age*, vol. 115, no 1, pp. 29-72.

Budassi et al. 2016 BUDASSI, M., PASSAROTTI, M. (2016). "Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon." In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. pp. 90-94.

Chastang 2007 Chastang, P. (2007). "Du locus au territorium. Quelques remarques sur l'évolution des catégories en usage dans le classement des cartulaires méridionaux au XIIe siècle." In *Annales du Midi : revue archéologique, historique et philologique de la France méridionale*, Tome 119, N°260, pp. 457-474.

Cohen et al. 2004 Cohen, W. W., and Sarawagi, S. (2004). "Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89-98.

Corrarati 1994 Corrarati, P. (1994). "Nomi, individui, famiglie a Milano nel secolo XI. Mélanges de l'Ecole française de Rome." *Moyen-Age*, 106(2), pp. 459-474.

Curran et al. 2003 Curran, J. R., Clark, S. (2003). "Language independent NER using a maximum entropy tagger." In *Proceedings of the seventh conference on Natural language learning* at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, p. 164-167.

Duby 1953 Duby, Georges (1953). *La société aux XIe et XIIe siècles dans la région mâconnaise. Bibliothèque Générale de l'Ecole Pratique des Hautes Etudes*, 6e section, Paris.

Durrell 2007 Durrell, M. (2007). "GerManC: a historical corpus of German 1650-1800: Full Research Report." *ESRC End of Award Report*, RES-000-22-1609.

Eger et al. 2015 Eger, S., Vor der Brück, T., Mehler, A. (2015). "Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods." In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH 2015), pp. 105-113.

Ehrmann et al. 2016 Ehrmann, Maud, et al. (2016). Diachronic "Evaluation of NER Systems on Old Newspapers." In *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016)). "Bochumer Linguistische Arbeitsberichte", p. 97-107.

Elson et al. 2010 Elson, D. K., Dames, N., McKeown, K. R. (2010). “Extracting social networks from literary fiction.” In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, p. 138-147.

Eltyeb et al. 2014 Eltyeb, S. and Salim, N. (2014). “Chemical named entities recognition: a review on approaches and applications.” *Journal of cheminformatics*, vol. 6, no 1, p. 17.

Erdmann et al. 2016 Erdmann, A., Brown, C., Joseph, B., Janse, M., Ajaka, P., Elsner, M., and de Marneffe, M. C. (2016). “Challenges and Solutions for Latin Named Entity Recognition.” *LT4DH 2016*, 85.

Frontini et al. 2016 Frontini, F., Brando, C., Riguet, M., Jacquot, C., and Jolivet, V. (2016). “Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data,” *MATLIT: Materialidades da Literatura*, 4(2), pp. 49-75.

Grover et al. 2008 Grover, C., Givon, S., Tobin, R., and Ball, J. (2008). “Named Entity Recognition for Digitised Historical Texts.” In *LREC*.

Guyotjeannin 1997 Guyotjeannin, Olivier (1997). “‘Penuria scriptorum’ : le mythe de l’anarchie documentaire dans la France du Nord (Xe-première moitié du XIe siècle),” *Bibliothèque de l'école des chartes*. Tome 155, livraison 1, pp. 11-44.

Hoffart et al. 2013 Hoffart, J., Suchanek, F. M., Berberich, K. and Weikum, G. (2013). “YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia Artificial Intelligence,” special issue on Wikipedia and Semi-Structured Resources.

Iogna-Prat et al. 2013 Iogna-Prat, D., et al. (2013), *Cluny : les moines et la société au premier âge féodal.* Presses universitaires de Rennes. Collection « Art et Société ».

Klein et al. 2014 Klein, E., Alex, B., Clifford, J. (2014). “Bootstrapping a historical commodities lexicon with SKOS and DBpedia.” In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH), pp. 13-21.

Lafferty et al. 2001 Lafferty, J., McCallum, A., and Pereira, F. (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” In *Proceedings of the eighteenth international conference on machine learning*, ICML, Vol. 1, pp. 282-289.

Lavergne et al. 2010 Lavergne, T., Cappé, O., and Yvon, F. (2010), “Practical very large scale CRFs.” In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 504-513.

Lehmann et al. 2013 Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, Sö. & Bizer, C. “DBpedia - A Large-scale, Multilingual Knowledge Base Extracted” from *Wikipedia Semantic Web Journal*, 2013

Li et al. 2016 Li, H., and Shi, J. (2016). “Linking Named Entity in a Question with DBpedia Knowledge Base.” In *Joint International Semantic Technology Conference*, Springer International Publishing, pp. 263-270.

Magnani 2002 Magnani, Eliana. (2002). “Le don au moyen âge.” *Revue du MAUSS*, no 1, pp. 309-322.

Mosallam et al. 2014 Mosallam, Y., Abi-Haidar, A., Ganascia, J. (2014). “Unsupervised named entity recognition and disambiguation: an application to old French journals.” In *Industrial Conference on Data Mining.* Springer, Cham, pp. 12-23.

Nadeau et al. 2007 Nadeau, D., Sekine, S. (2007). “A survey of named entity recognition and classification.” *Lingvisticae Investigationes*, vol. 30, no 1, pp. 3-26.

Neelakantan et al. 2015 Neelakantan, A., Collins, M. (2015). *Learning dictionaries for named entity recognition using minimal supervision.* arXiv preprint arXiv:1504.06650.

Nouvel et al. 2016 Nouvel, D., Ehrmann, M., Rosset, S. (2016). *Named Entities for Computational Linguistics.* ISTE, cognitive science series.

Passarotti 2014 Passarotti, Marco. (2014). “From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin.” In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH). pp. 100-109.

Plank 2016 Plank, B. (2016). *What to do about non-standard (or non-canonical) language in NLP.* arXiv preprint arXiv:1608.07836.

Pochampally et al. 2016 Pochampally, Y., Karlapalem, K., Yarrabelly, N. (2016), “Semi-Supervised Automatic Generation of Wikipedia Articles for Named Entities.” In Wiki@ ICWSM.

Rayson et al. 2017 Rayson, P., et al. (2017). “A deeply annotated testbed for geographical text analysis: The Corpus of Lake District Writing.” In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*. ACM, pp. 9-15.

Rizzo et al. 2011 Rizzo, G., Troncy, R. (2011). “Nerd: evaluating named entity recognition tools in the web of data.” In *Workshop on Web Scale Knowledge Extraction* (WEKEX’11).

Rosé 2007 Rosé, I. (2007). “Panorama de l’écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles).” *Bulletin du centre d’études médiévales d’Auxerre*, BUCEMA, (11).

Sopena 1996 Sopena, P. M. (1996). “L'anthroponymie de l'Espagne chrétienne entre le IXe et le XIIe siècle.” *L’anthroponymie document de l’histoire sociale des mondes méditerranéens médiévaux, Publications de l'École française de Rome*, 226(1), pp. 63-85.

Wallach 2004 Wallach, H. M. (2004), “Conditional random fields: An introduction.” *Technical Reports* (CIS), pp. 22.

Won et al. 2018 Won, M., Murrieta-Flores, P. and Martins, B. (2018) “Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora.” Front. *Digit. Humanit.* 5 :2. doi: 10.3389/fdigh.2018.00002

Zimmermann 2003 Zimmermann 2003 Zimmermann, M. (2003), *Écrire et lire en Catalogne : IXe-XIIe siècle (Vol. 1).* Casa de Velázquez, Madrid, pp.251-284.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.