Named entity recognition applied on a data base of Medieval Latin charters. The case of chartae burgundiae.

Sergio Torres Aguilar[1], Xavier Tannier[2], Pierre Chastang[3]
[1] DYPAC, Université Paris-Saclay (regester3@gmail.com)
[2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay (xavier.tannier@limsi.fr)
[3] DYPAC, Univ. Versailles St-Quentin, Université Paris-Saclay (pierre.chastang@uvsq.fr)

Abstract

The work on the named entity recognition (NER) in databases of historical texts has been placed among the most promising new ways to implement best recovery and managements tools for exploring mass data. In this paper, we describe the application processing NER through a modelling with CRF on an annotated database of Burgundy collection of charters from the tenth to thirteenth centuries. The aim is to generate a model for automatic recognition of named entities in historical sources. We discuss the nature of historical documents in the corpus and extraction of rules, and we expose adaptation to the processing algorithm and the most common problems encountered in Medio Latin texts using diplomatic formularies, which is an atypical case within the NER studies.

1 Introduction

In this paper we present the creation of an automatic recognition model of named entities on historical sources in Medio-Latin language. The benefits of NER applied to digitized editions of manuscripts are well known. The main hypothesis is that the control and management of indexing, through digital languages, of the names of people, places and institutions within vast databases could enrich results for information retrieval engines and provide unreleased data from edited sources. However, the production of an annotated corpus is a time-consuming and labor-intensive work. A research team can take months to annotate a corpus containing thousands of documents.

Our work uses an annotated database counting 5300 documents from editions of cartularies from Burgundy (10^{th}-13^{th} centuries) completed by the CBMA group*. On this basis, we intend to generate a model that can automate or semi-automate the recognition of entities in Medio-Latin language, used in most of the formal documentation written during the 10^{th}-15^{th} centuries. The model should be adaptable to different scriptural variants that developed according to the historical and institutional evolutions of writing practices. To accomplish this, we propose a statistical modeling linear chain based on CRF (Conditional Random Fields), a machine-learning method for labeling sequence. This method makes it possible to use a large number of features, from the words of a text to POS tagging, lemmas, suffix, punctuation, etc., for covering all aspects of a word inside a phrase context.

A model capable of automating entities recognition can offer, within a reasonable time, an enriched text with semantic information and structural level data. This will generate historical works that penetrate long-time realities poorly explored by now [1]. Moreover, the accelerated mass homogenization of textual data could favor the study of historical sources with renewed quantitative and statistical methods.

* This work uses the results obtained in the CBMA project (http://www.cbma-project.eu/), which aims to provide an open access database of the diplomatic sources produced during the Middle Ages in Burgundy and promotes studies on epistemological transformations of the research in Humanities generated by the digital tools. We want to thank and credit the network of researchers involved in CBMA project.

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

2 Previous work

Studies on the NER have a long history in other fields of textual analysis such as journalism and medicine [2] and most recently in social networks. In recent years, digital humanities have had among its objectives the recognition of named entities as one of the most promising fields for the discovery of new ways of exploration and new approaches to the huge masses of digitized documents. In the past five years, new tools for the exploration on huge masses of digitized data have been developed. Literary data bases as Gutemberg [3] or databases of digitized newspapers [4] have developed based-NER tools for exploring written phenomena. More specific works on the named entities detection can also be found in corpus of English parliamentary records of modern times [5] in Anglo-Occitan corpus of medieval times [6] or in bibliographic records, which also approaches semantic relationships [7].

3 Process

3.1. Corpus description

We use a database with nearly 19 thousand items obtained from 59 diplomatic editions from 300 cartularies and collections of charters, produced in Cluniac and Cistercian abbeys of Burgundy. Researchers responsible for the CBMA project (Chartae Burgundiae Medii Aevi) isolated a 5300-item corpus from the tenth to thirteenth centuries, produced in Cluniac abbeys, on which a manual annotation of named entities (personal names and place names) was performed. The arrangement of these data also follows a pattern which distributes in columns all about its identification

A cartulary is a volume in which the originals or transcripts of various kinds of documents, royal charters, privileges, judgments, notarial minutes, etc., are collected. These are formal documents that follow a model form, i.e. a stereotyped discursive structure on which names of people, places and organizations, dates, titles, etc. are added. All the documents are written in a medieval variant of Latin. Since the texts come from diplomatic editions, they include elements that are not usually found in the original, such as punctuation, capitalization and development of abbreviations.

3.2. Processing corpus

Processing Latin requires to compile various observations related to the morphology and syntax of the language regarding the occurrences of named entities. As Latin is an inflected language, the function of the word is usually contained in the termination. Personal names usually appear in the nominative and accusative cases (-o, -us, -um ending). The genitive in the Medio Latin variant is problematic (-is, -isis, -orum ending) because it presents in some cases a name and a place name without separation, e.g.: *Armundi* (name) *Vianensis* (place) *archiepiscopi*. There is also a notable shortage of ablative and dative cases. Flexion also causes a restricted use of prepositions, and, in addition, the order of the word in the sentence reduces its importance. Therefore, the information about the lemma and the suffix is crucial to work with NER. Besides, the attenuation and corruption of the rules of classical Latin texts cause a series of irregularities such as the abuse of enclitics, lexical redundancies and significant instability in the writing and spelling of names.

However, in syntactic terms, the context of appearance of simple entities in Medieval Latin is not very different from the context in Romance languages. Important characteristics include, but are not limited to:

- *Classifying type words*: villa, terra, locum, mansum, castrum
- *Titles and functions:* Sanctus, dominus, presbiter, comes, rex, episcopus
- *Prepositions and affixes*: ad Artulfo, in villa Verziaco, per manum Aymini, pro anima Fromaldi,

Albeit compound entities are very original of the Medio Latin language:

- *Hagiotoponyms*: ecclesia in onore Sancti Andree; terra Sancti Petri; partes Sancti Petri campi
- *Two-element anthroponyms*: Matheus de Sosiaco; Hugo Gregarii; Odelinus filius Aginaldi; Petrus dicitur Grossus
- *Appositions*: Willebertus nomine, Regnante Lothario Rege; Quod Guido cantor

In many cases, it is the meaning of the word and not its morphology which determines its function (E.g. In alio [terre Adalrado (person)] place). In other cases, there is a confusion between personal names, places and institutions, for example in the so-called *donatio pro anima* (E.g. Donamus a sancti Pauli et Sancti Petri Cloniacense in pago Vianense), where the donation of land or property may be made under the invocation of a saint (person) and it becomes part of a land dedicated to the saint (place), but the donation can be administered by a church dedicated to this saint (institution).

Entities related to names of saints can also lead to confusion because their appearances may have different contexts of use: festivities, biblical references, allegories, invocations or donations. Most of these entities do not offer greater utility to our model and we have not considered it in the annotation.

A very problematic case is that of overlapping or nested entities: a person's name including or appearing next to a place name or within an institutional name. Nesting levels are usually two, but can get to the third level. For example:

This phenomenon involves the addition of two or more columns of labels to certain tokens, which complicates the treatment of the data, because most machine learning classifier are not designed to attributes more than one class to each instance. The loss of information at this level can be critical for further work because this phenomenon reveals the genesis and evolution of the compound name and preserves information about familiar and social relationships [8]; all of this data serves for attributing an individual ID, which is the important one for History, to named entities, which is what we retrieve.

On the other hand, the initial corpus only contains marks of recognition for names and places, which is a serious problem when an entity that clearly should be classified as an institution appears. The use of classic trinomial of ENAMEX: person, place, institution, would involve a long manual extension of the original corpus. At the same time, for first experiments, we want a model applied only to individuals and not legal and administrative entities. For this reason we conserved the original annotation of about 95% of the institutional entities as places.

In general, the recognition of entities in Medieval Latin does not only concern the lexical and morphological properties of the word, but also its semantic and historical context. This entire series of accidents eventually lead to a long work of manual validation of lists extracted from the corpus containing the problematic entities. Correcting these lists provides an update of the current annotation guidelines to lead to our gold-standard corpus.

4 Method

The corpus was transformed into a 7 columns format with lexical, morphological and semantic information: TOKEN, POS, LEMMA, CASE, SUFFIX, NAME_Entity, GEO_Entity. The first three columns were obtained from a variant of TreeTagger for the Medio Latin language created by the OMNIA group in 2013[†]. The two following columns add information about the capital letters in the word and give the suffix (last three letters) of each word. For columns containing named entities we used BIO labels, which are very useful to determine categories and boundaries of named entities through

[†] http://www.glossaria.eu/treetagger/

a very simple format: B-X, I-X and O for representing the beginning, continuation or absence, respectively, of named entities.

Once transformed the textual corpus to CRF format, we started a machine-learning-based recognition method. The total number of documents was later divided semi-randomly in well distinct parts into two sections: a major section of 4300 documents for the development corpus and a minor section of 1000 documents for test corpus. Later, the development corpus was split in the training corpus set of almost 4000 documents (> 1 million words) for training the algorithm, and the rest for the development test-set, which would reduce the degree of extrapolation model.

In addition we create a pattern that determines word by word the rules of observation for the document and the relevant combinations of information both in lines and columns, i.e., the lemma, the entity and the phrase itself. Combining unigrams and bigrams, we wrote a pattern of 26 unigrams with an extended sequence of two positions ahead and two positions behind in the token column for each line (see Table 1).

For this work, we have annotated in the same format provided for the database and we used Wapiti[‡] [9] a toolkit for labelling sequences developed by LIMSI-CNRS with standard options for the work with CRF linear-chain: L-BFGS algorithm that rendered better results than others such as BCD and RPROP+ and defaults values for L1 and L2 regularization (0.5 and 0.0001 respectively).

TOKEN	POS	LEMMA	CASE	SUFIX	ENTITY	ENTITY
Quod	CON	quod	UPPER	uod		
ego %x[2,0]	PRO	Ego	LOWER	ego		
Hugo%x[1,0]	NAM	-	UPPER	ugo	B-PERS	
de %x[0,0]	PRE [0,1]	de [0,2]	LOWER [0,3]	de [0,4]	I-PERS[0,5]	[0,6]
Berziaco %x[-1,0]	NAM	-	UPPER	aco	I-PERS	B-LOC
perpendens %x[-2,0]	VBE	perpendeo	LOWER	ens		
,	PON	,	LOWER			

Table 1: Sequence labels and relevant positions from pattern for each line in 7 columns format.

5 Results

The performance with all this data was remarkable, exceeding, 96% F-measure on test data for the person's names in beginning of entities and 92% F-measure for place names. The number has decreased for identification on the rest of entity in about 88% F-measure for person entities and more discrete result in the case of locations, 80%. In general, model precision is very close to model recall in person names and locations, but, in the second case the distance from B-LOC to I-LOC is about 10% less.

Person name	Precision	Recall	F1
B-PERS	0.95	0.96	0.96
I-PERS	0.88	0.92	0.90
Location name			
B-LOC	0.91	0.93	0.92
I-LOC	0.81	0.80	0.80

Table 2: Best current ratio recognition

Our model is very strong to distinguish single entities and entities in combination with master words preceding the apparition of an entity like *villa, terra, ecclesia, mansus,* for the places and *ego, filius, uxor, abbas, frater*, etc for the names. Moreover, our model is strong to recognize an easy combination of name entity and to place the entity under the format *name + de + place* where the origin of the format name and last name is. But the model is less able (but still strong) to capture boundaries of entities into

[‡] https://wapiti.limsi.fr/

long place names, especially when the composition is a combination (or overlap) of institutional and geographical names.

The rate of error in manual annotation with this kind of entities is stronger because of the ambiguity, and in an examination line by line of the result we noticed that a very important quantity of errors in our results is linked to errors in the annotation of our standard-gold corpus. The high rate of recognition is surely connected with the specific nature of the document where are a textual formulaic patterns.

6 Future work

Three immediate improvements to the system are being planned:

i. We will increase further iterations of the model with corpus of increasingly less voluminous training to reach the most balanced outcome between the extent of the training corpus and the recognition rates.

ii. Besides, we will test the robustness of the model with 400 new documents from different periods from the large corpus that we have annotated by hand as well as on other documents beyond our corpus. We will test the model on same type of documents: charters and cartularies, and on different document types, such as Latin chronicles or administrative documents.

iii. A likely new review of the quality of the gold-standard corpus in order to correct systematically some manual errors in the annotation.

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

References

1. Moretti, F. (2013). *Distant reading*. Verso Books.
2. Bodnari, A., Deleger, L., Lavergne, T., Neveol, A., & Zweigenbaum, P. (2013, September). *A Supervised Named-Entity Extraction System for Medical Text*. InCLEF.
3. Brooke, J., Hammond, (2015). Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. 4th Workshop on Computational Linguistics for Literature, 42-47.
4. Mac Kim, S., & Cassidy, S. (2015). *Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers*. In Australasian Language Technology Association Workshop, 57-60
5. Grover, C., Givon, S., Tobin, R., & Ball, J. (2008, May). *Named Entity Recognition for Digitized Historical Texts*. In LREC.
6. Scrivner, O., & Kübler, S. (2015, June). *Tools for digital humanities: Enabling access to the old occitan romance of flamenca*. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, 1-11.
7. Byrne, K. (2007, September). *Nested named entity recognition in historical archive text*. In Semantic Computing, 2007. ICSC 2007. International Conference on IEEE, 589-596
8. Beck, P. (1996). *Anthroponymie et parenté*. Collection de l'Ecole française de Rome, 226, 495-496.
9. Lavergne, T., Cappé, O., & Yvon, F. (2010, July). *Practical very large scale CRFs*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 504-513