Extracting Information from Medieval Notarial deeds*

Charlene Ellul[1], Joel Azzopardi[1] and Charlie Abela[1]

University of Malta, Malta
charlene.ellul@um.edu.mt joel.azzopardi@um.edu.mt charlie.abela@um.edu.mt

Abstract. The Notarial Archives in Valletta houses a collection of Latin Notarial deeds that has not been exploited yet. In this paper, Machine Learning techniques are proposed and implemented to extract entities such as people, place names, dates, deed types and keywords from these historical texts. Both supervised and unsupervised techniques are considered and compared with baseline models. Experimental results on a subset of these documents are already showing results that outperform the baselines for Latin text such as those from CLTK. Evaluation was carried out using indexes of four published Notarial Registers.

Keywords: Named Entity Recognition · Keyphrase Extraction · Text Classification · Latin Text · Historical Texts

1 Introduction

Archives around the world are a source of hidden information. One of these archival gems is found at the Notarial Archives in Valletta, Malta and houses around 20,000 notarial deeds dating back to the 13th century. It is typically the case that archives publish high quality images and metadata of the structure of historical documents, but the content itself is not exposed in a meaningful way to aid historians. In literature, some researchers [1] dedicated their efforts to mine data from medieval Latin documents. The extraction of named entities such as dates, places and people can be used to aid historical research in areas such as genealogy and toponymy.

Most of the notarial deeds found in the Valletta archives fall under categories such as wills, dowry and transfer of land. Although some notaries used to write the deed type, this was not a requirement. Thus, text classification can be used to maximize the use of the remaining words to predict the presented deed type. Automatic keyphrase extraction can be used to express a document as a set of keywords/keyphrases[2]. A notarial deed can be represented in a similar way to shed light on its content and avoid unnecessary handling whilst also reducing the time required for archival researchers and enthusiasts to find what they are looking for.

* This work is partially funded by project E-18LO28-01 as part of the collaboration between the Notarial Archives in Valletta and the University of Malta.

In our research we use four Latin notarial transcribed registers entitled *'Documentary Sources of Maltese History'* and compiled by Professor Stanley Fiorini[3]. These are the only existent transcribed documents from the collection dating back to the 15th century with 981 deeds. Our main goal is to extract entities such as dates, people, places, deed types and keyphrases. Annotating a large corpus of data requires expertise and time, fortunately, however, these publications include indexes for place names, persons and subjects which could be used to annotate the deeds and also for evaluation.

In the rest of the paper, we discuss some related work in Section 2 and present the adopted methodology in the following section. In Section 3 we present some initial evaluation which is followed by some future work.

2 Related Work

Latin poses a great challenge for Named Entity Recognition (NER). Annotated corpora that can be used for training are scarce and most of them focus on classical texts. Conditional Random Fields (CRF) have been used successfully for training Latin models as in Aguilar et al. [1] who applied CRF on a database of Burgundy cartularies which were manually annotated. Text classification techniques based on supervised machine learning have also been used in the context of different languages[4].

Both supervised and unsupervised models for Keyphrase detection and extraction have been used successfully[2]. A keyphrase extraction model is usually based on a list of extracted candidate words and some heuristic such as stopword removal through which candidate keywords are filtered out. RAKE[1], TextRank[2] and TF-IDF are three popular unsupervised approaches that have been applied on generic languages [2]. A more domain-specific keyphrase extraction method was developed by Witten et al. [5] who designed the KEA supervised algorithm. Candidate keyphrases (up to 3 words) are filtered before computing TF-IDF and the distance from the start of document for each candidate as features and then fed into a Naive Bayes Model. CRF were used by Zhang et al. [6] using a number of features among which are length of word, POS tag, previous words, next words and TF-IDF. This was tested on a Chinese text and yielded the best F1 score compared to SVM and other baseline models.

Methodology

We used the indexes found in [3] to annotate the text with people's names and places for the NER, and keyphrases for keyphrase extraction. Typically a keyphrase index has the following form, *Coquine domus/domuncula 226, 241-242, 396*, with the term and the deeds containing the term. There exists however an electronic version of a single index, while the other copies are available as hard

[1] https://github.com/fabianvf/python-rake
[2] https://github.com/davidadamojr/TextRank

copies. Furthermore, there is no index available for dates. A dictionary of all possible mentioned entities was compiled using the Ratcliff-Obershelp distance algorithm[3] to annotate the text with the relevant tag to be used for evaluation.

Dates are presented using indictions for the year and thus we had to work out the indiction cycle of each act. Notaries tend to use shorthand when writing dates such as *eodem* (same date as before) and *penultimo* (day before the last). For this reason a rule based entity extraction was implemented to convert the dates to modern dates. Extraction of person and place entities was performed using a trained CRF model based on suffixes, uppercase, title, digit, POS, lemma, next and previous words. The POS and lemma tags were derived using Schmidt's treetagger[4] using parameters for Latin to improve accuracy. We used the Fiorini's register [3] to train and test our model which was then compared with existing libraries such as the Classical Language Toolkit (CLTK) [5], Spacy (multilingual model)[6] and Stanford's NER (Spanish Model - Latin derivative)[7].

Deeds can have a variety of categories and these were generalized for text classification. In total there are 981 deeds and 79 different categories with an average of 12 deeds per category. Some of the categories include only one deed, making the training set highly imbalanced. Different feature vectors were tried out including count vectors, word level TF-IDF, n-grams and character level vectors. Different models were trained for deed classification using Naive Bayes, Linear Classifier, SVM and Random Forest. The model with the highest accuracy was saved.

The index of keyphrases was used to annotate the corpus for keyphrase extraction. Lemmas were used for comparisons as Latin often uses declensions. The index was merged with the deed text using exact, lemma and stem matches. A list of annotated/non-annotated words was kept for each deed to be used for evaluation. Generic unsupervised approaches were used for keyphrase extraction including TextRank, RAKE and TF-IDF, however these yielded unsatisfactory results with RAKE giving the best results as shown in Table 1. We then used a variant of the supervised approach presented by [5] called KEA which due to its candidate phrases filtering did not yield good results. A CFR algorithm was implemented using the same technique to extract entities for people and places, with the addition of TF-IDF and distance features giving the best results as shown in Table 1.

3 Evaluation

The results achieved for the extracted entities are already very promising. Both NER and keyphrase extraction were done using 100 acts (indexes of other registers are yet to be used for annotations) with a 70%-30% split (results in Table 1).

[3] https://docs.python.org/2/library/difflib.html
[4] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[5] http://docs.cltk.org/en/latest/index.html
[6] https://spacy.io/
[7] https://nlp.stanford.edu/software/CRF-NER.shtml

Since we used a domain specific corpus, supervised models gave better results. The dataset was highly imbalanced and the text classification was performed on the whole corpus of 981 records with a 75%-25% split. A Linear classifier was used that leveraged on CLTK's stopword list and the count vector features giving an accuracy of 72%. The removal of stopwords improved slightly the achieved results across all trained models.

Table 1. Name Entity Recognition and Keyphrase extraction results

Purpose	Method	Precision	Recall	F1 score
NER People/Places	CLTK	0.339	0.113	0.17
NER People/Places	Spacy with multilingual model	0.414	0.922	0.572
NER People/Places	Stanford NER with Spanish model	0.152	0.947	0.263
NER People/Places	Conditional Random Fields Model	0.956	0.957	0.956
Keyphrase extraction	RAKE with CLTK and Voyant tools stop words[1]	0.257	0.15	0.189
Purpose	Method	F1 score		
		O-KEY	B-KEY	I-KEY
Keyphrase extraction	Conditional Random Fields	0.982	0.751	0.465

[1] https://github.com/aurelberra/stopwords

4 Future Work

In this paper, we presented our initial research on extracting entities and keyphrases from historical Latin texts. The results are very encouraging even though the datasets are fairly small. We plan to digitize the indexes of the other registers in Fiorini's collection so that we can train the models with more data. We will furthermore be using the extracted information to create a knowledge graph for the Notarial Archives.

References

1. S. T. Aguilar, X. Tannier, and P. Chastang, *Named entity recognition applied on a database of Medieval Latin charters. The case of chartae burgundiae.* M. Dring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch, 2016, pp. 67–71.
2. K. Saidul Hasan and V. Ng, *Automatic Keyphrase Extraction: A Survey of the State of the Art.* Association for Computational Linguistics, 2014, pp. 1262–1273.
3. S. Fiorini, *Documentary Sources of Maltese History Part I Notarial Documents No 3 Notary Paulo Bonello, Notary Giacomo Zabbara*, 1st ed. University of Malta, 2005.
4. A. Al-Thubaity, N. Abanumay, S. Al-Jerayyed, A. Alrukban, and Z. Mannaa, "The effect of combining different feature selection methods on arabic text classification," in *2013 14th IEEE/ACIS, SNPD*, 2013, pp. 211–216.
5. I. H Witten, G. W Paynter, E. Frank, C. Gutwin, and C. G Nevill-Manning, "Kea: Practical automatic keyphrase extraction," 1999.
6. C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems*, pp. 1169–1180, 2008.