LOCALE: A rule-based location named-entity recognition method for Latin text[*]

Ivona Milanova[1], Jurij Šilc[2], Miha Seručnik[3], Tome Eftimov[2,4], and Hristijan Gjoreski[1]

[1] Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius - University, Skopje, North Macedonia
ivonamilanova221@gmail.com, hristijang@feit.ukim.edu.mk
[2] Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia
jurij.silc@ijs.si, tome.eftimov@ijs.si
[3] Milko Kos Historical Institute, Research Centre of the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia
miha.serucnik@zrc-sazu.si
[4] Stanford University, Palo Alto, California, USA
teftimov@stanford.edu

Abstract

For creation of digital textual corpora of preserved historical sources, automatic or semi-automatic extraction of specific types of information is becoming a requested tool for many researchers active in the field of digital humanities. With such tools, the efforts in digitization and semantic annotation will be greatly aided. For this reason, we propose a rule-based named-entity recognition system that can be used for location extraction from Latin text (so-call LOCALE). It is based on a set of computational linguistics rules for Latin language. Experimental results obtained on a set of 100 documents, which were further manually evaluated by human experts, showed that very promising results are achieved.

1 Introduction

Historiography engages with a very wide variety of subjects, as long as they are connected to past human activity in some way. The traditional and still central source for the study of history has been and remains the written text. Historian's (textual) sources may vary widely in variety and quantity, but in general terms it is true the further back in time one reaches, the more scarce the sources.

Location names (settlements, castles, churches, fallows, waters, provinces, etc.) represent a specific genus of information commonly contained in all (or almost all) written sources. They, first and foremost, attest the existence of a given settled place at the time of creation of the text in which they appear. Furthermore, several social phenomenon may be inferred from location names such as the extent of political entities, medieval manors, but also enables insights into historical demography in the sense of the spreading of population over territories and so on.

Joined with personal names location names give the possibility to map the life paths of individuals in space and time, as well as to create models of interpersonal or other more abstract connections for a group of people mentioned in the sources. Such models may be built, depending on the type of sources utilized, for nobility or high clergy as well as for merchants or peasants. To sum, studies in social, economic, political, demographic, military, and many more aspects of history deal with topographic data in historical sources. Of course, other disciplines that are counted among the humanities may also make use of this sort of information.

Through creation of digital textual corpora of preserved historical sources automatic or semi-automatic extraction of specific types of information is becoming a requested tool for many researchers active in the field of digital humanities. With such tools, the efforts in digitization and semantic annotation will be greatly aided.

The long-term goal of the research will be to produce a thorough and reliable digital (interactive) version of historical topography of the entire territory of the present-day Republic of Slovenia. The first step in this objective has already been successfully achieved in 2015 with the project 'Slovenian place-names in time and space', which covered the territory of the historical province of Carniola [10]. The Web page is available at http://topografija.zrc-sazu.si/.

The reminder of the paper is organized as follow: Section 2 provides the related work, Section 3 presents a rule-based system for location extraction from Latin text, Section 4 explains the data used for evaluation together with the experimental results, and finally Section 5 concludes the paper and provides directions for future work.

2 Related Work

In the past few years, with a fast digitization of historical textual sources [11], the extraction of locations from historical text has become important for Spatial Humanities. One way to extract the locations from free text is to apply information extraction (IE), which is a task of automatically extracting information from unstructured (i.e., textual) data [3]. This task involves applying natural language processing (NLP) techniques or applying methods that can work with human natural languages [2]. After applying IE method, the result consists of predefined concepts (i.e., entities) of interest represented in a structured way. The information to be extracted from text is defined by users, and it is related to some specific domain.

Named-entity recognition (NER) is a sub-task of IE, which aims to determine and identify words or phrases in text into predefined classes that describe the entities of a given domain [13]. There exist several types of NER methods: terminological-driven [12], rule-based [4], corpus-based [16], active learning (AL) methods [6], and deep neural networks (DNNs) methods [17].

Terminology-driven NER methods [12] are the simplest ones. In their case, text phrases are matched with entities of a given domain that have already existed in dictionaries. Usually, to have better performance, the marching is not searching for perfect matches, but some heuristics, such as the generation of words that occur in entity mentions, generating permutations of words in concept synonyms, solving disambiguation problem, etc, are applied. The weakness of such methods is that they can only extract the entities that are part from the used dictionary(ies). *Rule-based* NER methods [4] are another alternative, where regular expressions, which combine information from dictionaries and the characteristics of the entities, are used in the extraction process. The main weakness of these methods is the manual construction of the rules, which can be a time-consuming task depending from the domain. *Corpus-based* NER methods [16] are the most commonly used. They use annotated corpus provided by domain experts, which is further utilized with machine learning (ML) algorithms to predict the entities' classes. The benefit of these methods is that they are less affected by dictionaries and manually created rules, but the weakness is the availability of an annotated corpus for the domain. To create an annotate corpus for a new domain is a time consuming task and requires huge effort by domain experts. One way to minimize the annotation cost is to apply *active learning* [6], which is an iterative supervised learning, where an algorithm is able to interactively query the user to obtain the desired outputs at new data points. Because corpus-based NER methods are based on costly handcrafted features to train NER model, recently a lot of work is done on NER based on *deep neural networks* (DNNs) [17]. The benefit of using them is that they do not need good handwritten features, but the weakness is that they typically require large amounts of annotated data.

Despite the large amount of research done in NER for other domains, there are several studies that are trying to address NER for historical text. In most cases, corpus-based NER methods are applied, which also depend form the annotated historical corpus. Working with historical corpus additional challenges should be considered, such as, language changes over time, spelling variations, transliterations, and sources written in multiple languages among others. In [16], the authors evaluated five classical corpus-based NERs used for location extraction from two historical datasets and combined them through an ensemble of majority vote. The datasets used for evaluation are the Mary Hamilton Papers [14], written in modern English, and the Samuel Hartlib collection [8], written in early-modern English. The evaluation results showed that individual performance of each NER is corpus dependent. To improve corpus-based NER, in [6], the authors proposed an active learning solution.

From other side, the collection of historical Latin text is still relatively sparse compared to English. In [5], the authors presented the first annotated corpus for NER in Latin text and a corpus-based NER that use language-specific feature set for training the model. Also, in [1], the authors presented a corpus-based NER based on conditional random field (CRF) [15] using an annotated corpus of Burgundy collection of charters from the tenth to thirteenth centuries.

3 A Rule-based Location Named-entity Recognition for Latin Text

As we have already mentioned, working with historical corpora, especially Latin text, is challenging because language changes over time, spelling variations, and transliterations. Since the performance of corpus-based NERs is dependent from the quality of an annotated corpus, and having an annotated corpus is a time-consuming task, we propose a rule-base location named-entity recognition method, called LOCALE (ru**L**e-based l**OC**ation n**A**med-entity recognition method for **L**atin t**E**xt), where the method does not using training annotated data, but a small set of manually created rules. The goal of this study is to find if such a method can also provide promising results that can be used for location extraction from Latin text.

3.1 LOCALE pipeline

Since we are lack of annotated corpus, to extract location from Latin text, we propose a rule-based named-entity recognition method, called LOCALE. It is based on a set (i.e., grammar) of computational linguistics rules specific for the Latin language. We consider that the Latin language has well defined rules for presenting locations with a small number of exceptions.

The Latin language is an inflected language, which means that the type of the word can be determined from its suffix. There are several examples when the suffix is related with location. For example, the suffix *-ensis* is an adjectival suffix meaning ‘originating in’ and it is used in modern Latin scientific coinages, especially derivatives of place names. Another suffix that can help in recognizing locations is the suffix *-ensem* which is the accusative singular of *-ensis*, as well as the dative variant, *-ensi* and the ablative *-ense*. The plural form of the suffix *-ensis* is *-enses* which is also used to indicate locations for example *Aquilegiensis* which derives from

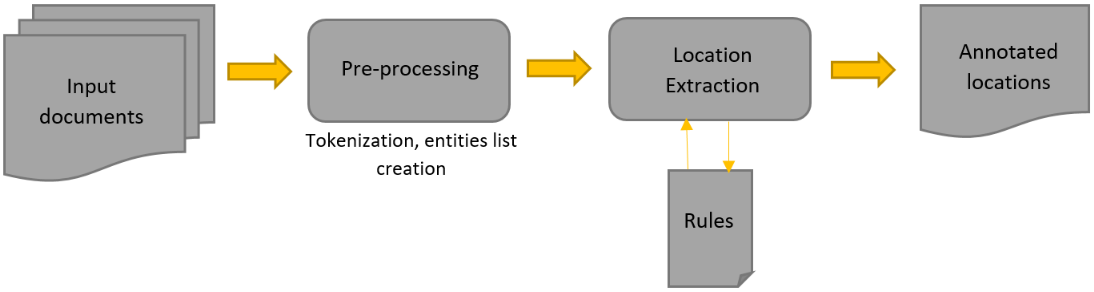

Figure 1: The LOCALE pipeline for location extraction from Latin text.

Aquilegia. The genitive and the accusative forms of the plural form of the suffix are -*ensium* and -*enses*, respectively. The suffix -*anus* has a similar meaning and it indicates a relationship of position, possession, or origin. The accusative singular form -*anum* and plural masculine form -*anarum* are also very commonly used. The suffix -*ius* and its dative -*iis* and accusative -*iam*/-*ium* forms are also used to form adjectives from nouns and indicates 'belonging to'. The -*ianus* suffix is frequently used in adjectives formed from proper names. Having a knowledge of these suffixes and their meaning can bring to a conclusion that the lemma and the suffix is crucial to work with NER.

Except for suffixes there are also several characteristics by which we can determine a location. Some of these characteristics are:

- Classifying type words: *villa*, *terra*, *mansum*.
- Geographical entities: *flumen*, *monte*.
- Titles and functions: *imperator*, *rex*, *episcopus*, *sancte*.
- Prepositions: *de*, *in*, *iuxta*, *ad*, *inter*.

Also, there are several issues that should be considered that can lead to a false positive location recognition. One such example are entities related to saints' names. This happens since their appearance may have different context, in some cases it signifies personal name and in some cases a location. For example in the phrase '... *Ego Iohannes sancti Nicolai in Carcere Iulianus* ...', *sancti Nicolai* is a location and in the case '... *sancti Benedicti Aquilegensis* ...', *sancti Benedicti* is a personal name.

Another issue are some of the suffixes that in the the majority of cases are an indication of a location. However, these suffixes in some cases can be found as an ending of a personal name. For example the suffix -*ium* indicates a location in the example '*versus Rosacium*' and a personal name in '*prefatum Wernerium*'.

LOCALE consists of two steps:

- Each document (i.e., Latin text) is going under **pre-processing** by applying tokenization and generating list of entities that are usually followed by a location or come right after the location.
- The list of tokenized words are checked by the set of computational linguistics rules, which define LOCALE grammar, in order to **extract** the location entities.

The LOCALE pipeline is presented in Figure 1.

The first step, that is the pre-processing, consists of a creation of a static list of entities by which we can determine a location like geographical entities (e.g., *rivus*, *monte*), titles and functions (e.g., *imperator*, *episcopus*) etc. This list was further used as a look-up table for checking the previous and following word of each individual token. The next pre-processing step is tokenization where we split the input document into individual tokens/words which are used as the basic input elements for the next step which is application of rules.

The LOCALE grammar consists of four rules:

- The first rule check each word if it stars with a capital letter. This is the first condition to be fulfilled for a further rule check to be applied on the specific word. This comes from the fact that every location in Latin starts with a capital letter.

- The second rule extracts words that are after the adverbs like *de*, *in*, *iuxta*, *versus*, *ad*, *inter* and so on. This comes from the fact that the adverbs in most of the cases are followed by a location. One problem that appears here is the listing of multiple locations after one of those adverbs, for example, '... *iuxta Natissam, Stanowiza, Boriana, Potoch, Creda* ...'. For this reason, we extended this rule to take into account multiple locations that are after an adverb. An exception of this rule was also applied. This rule covered the cases of listing people that belong to a specific location rather than locations, like '... *domini Bertoldi senioris de Rosaciis, Henrici, Rantolfi* ...' where *Henrici* and *Rantolfi* are personal names.

- The third rule is related to words coming after and before the word. As mentioned above geographical locations, titles and functions are a good indicator that the following or previous word is a location. For this purpose we created a dictionary lookup list that consists of entities belonging to the title and function categories as well their derivations. For example, *ecclesiam* and *ecclesias* present the accusative case in plural and singular of the word *ecclesia*, respectively.

- Finally, all other words that are not recognized by the three rules, are additionally checked for the abovementioned suffixes that are related to locations.

The LOCALE source code is publicly available at https://github.com/Ivona221/LOCALE.

4 Evaluation

In this section, we explain the data set used for evaluation of our proposed method, together with the obtained results.

4.1 Data

Historical topography deals with the identification of former location names that are found in various written historical records of the past. Since our group at the Milko Kos Historical Institute, Ljubljana, is highly involved in researching historical geographical locations (i.e., *historical topography*, which deals with the identification of former place names), to test the performance of the proposed method, we selected a data set that consists of medieval historical charters written and published at the territory of the Republic of Slovenia and its surrounding area.

Early documents produced in the present-day Slovenian territory were written mainly in Latin [9]. Therefore, as data set we took 100 documents (from historical critical edition) of the Abbey of Rosazzo from twelfth and thirteenth centuries (between 1132 and 1249) [7]. The monastery is located in the Colli Orientali of Friuli, around ten kilometres west from the Slovenian border. We limited on a small number of testing documents since the results were manually evaluated by domain experts from the Milko Kos Historical Institute, since we did not have an annotated corpus.

The test data set is publicly available https://github.com/Ivona221/LOCALE.

4.2 Results and Discussion

After applying LOCALE on 100 documents, the result consists of location entities extracted from each document. Because we do not have an annotated corpus (i.e., the ground truth), the results were manually evaluated by human experts. In our case, we have one positive class, which is the location.

Going through the extracted entities, the human experts classified each of them as a true positive (TP), which means that the extracted entity from the method belongs to the positive class (i.e., a location entity is truly recognized), or a false positive (FP), which means that the extracted entity is incorrectly classified as location entity (i.e., the entity is not a location.) Additionally going through the Latin text, they also report the location entities that were not recognized by the method, or these entities are classified as false negatives (FNs) (i.e., the entity that is incorrectly classified that does not belong to the positive class.) We should point here, that the results provided by both human experts was the same. The number of TPs, FPs, and FNs is presented in Table 1.

Table 1: Manually evaluation.

True Positive (TP)	302
False Positive (FP)	26
False Negative (FN)	29

Using the results reported in Table 1, the evaluation metrics for F_1 score, precision, and recall, are presented in Table 2. Precision gives the percentage of the results that are relevant, while the recall provides the percentage of total relevant results correctly recognized by your method. The F_1 score is a measure of accuracy, which is the harmonic mean of precision and recall. Higher values are preferable for all three evaluation metric.

Table 2: Evaluation metrics.

F_1 Score	Precision	Recall
0.9164	0.9207	0.9123

In our evaluation, the last rule related to the suffixes provides the most of the annotated locations. The most common suffixes that were present in our corpus were *-ensis*, *-anus* and *-us* and their derivations in genitive, dative, accusative and ablative, as well as their plural forms.

Looking at the evaluation results, we can concluded that the first attempt of extraction location entities from Latin text using a rule-based system provides very promising results. However, since our evaluation corpus consists of a 100 documents, we are planning to extend this work in future by:

- Adding more number of documents that will be used for evaluation;

- The documents will be extracted from larger number of charters in order to have more variability in the Latin text;

- Comparing LOCALE with the corpus-based NER method presented in [6] using their annotated corpus and using our own constructed corpus using charters written and published at the territory of the Republic of Slovenia and its surrounding area.

- Sensitivity analysis of the LOCALE with regard to Mediaeval Latin variation over time and grammar.

- Creating an annotated corpus that can be used for training corpus-based NERs.

5 Conclusions

To extract locations from Latin text, we proposed a rule-based named-entity recognition method, called LOCALE. The method is based on a small number of computational linguistics rules that are specific for the Latin text. The evaluation was performed using early documents produced in the Slovenian territory from twelfth and thirteenth centuries. Experimental results obtained using a 100 documents, which were further evaluated by human experts, showed that promising results can be achieved resulting in 0.92 for precision, 0.91 for recall, and 0.91 for F_1 score.

6 Acknowledgments

This work was partially supported by the project Toponomastical heritage of Primorska Region, which received funding from the Slovenian Research Agency (grant number L7-9424). J.Š. and T.E. acknowledge the financial support from the Slovenian Research Agency (research core funding number P2-0098).

References

[1] Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *HistoInformatics@ DH*, pages 67–71, 2016.

[2] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[3] Jim Cowie and Yorick Wilks. Information extraction. *Handbook of Natural Language Processing*, 56:57, 2000.

[4] Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488, 2017.

[5] Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. Challenges and solutions for latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, 2016.

[6] Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, 2019.

[7] Reinhard Härtel and Cesare Scalon. *Urkunden und Memorialquellen zur älteren Geschichte des Klosters Rosazzo.* Österreichische Akademie der Wissenschaften, 2019.

[8] Samuel Hartlib, John Dury, et al. *Samuel Hartlib and the advancement of learning.* Cambridge University Press, 1970.

[9] Dušan Kos. *Document, writing, writer. A contribution to the history of the Carniolian documents up to 1300.* Zgodovinski arhiv Ljubljana, 1994.

[10] Miha Kosi, Matjaž Bizjak, Miha Seručnik, and Jurij Šilc. *Historična topografija Kranjske (do leta 1500).* Slovenska historična topografija, 1. Založba ZRC, Ljubljana, 2016.

[11] Gerhard Jan Nauta and Wietske van den Heuvel. Survey report on digitisation in european cultural heritage institutions 2015. *Mode of access http://www. den. nl/art/uploads/files/Publicaties/ENUMERATE_Report_Core_Survey_3_2015. pdf.[In English]*, pages 2013–2015, 2015.

[12] Goran Nenadic, Irena Spasic, and Sophia Ananiadou. Terminology-driven mining of biomedical literature. *Bioinformatics*, 19(8):938–943, 2003.

[13] Sune Pletscher-Frankild and Lars Juhl Jensen. Design, implementation, and operation of a rapid, robust named entity recognition web service. *Journal of cheminformatics*, 11(1):19, 2019.

[14] Amy Prendergast. *Literary salons across Britain and Ireland in the long eighteenth century.* Springer, 2015.

[15] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

[16] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. Eensemble named entity recognition (NER): evaluating NER tools in the identification of place names in historical. *Frontiers in Digital Humanities*, 5:2, 2018.

[17] Yingwei Xin, Jean-David Ruvini, and Ethan J Hart. Deep hybrid neural network for named entity recognition, February 28 2019. US Patent App. 15/692,392.