APPLICATIONS OF ONTOLOGIES AND NEURAL MODEL IN THE PROCESSING OF HISTORICAL DOCUMENTS

DANIELA-MARIA CRISTEA, DANA LUPSA, ALIN MIHIS, BOGDAN-GABRIEL TROFIN

Abstract. This paper addresses the need to use natural language processing techniques in historical document analysis to aid historians and scientists by arranging metadata so that they can check for appropriate considerations among all processed documents. Through various elements such as science utility, accessibility to late medieval documents to research events or 14-16 century private history, we find its use for academic purposes and the benefit of the society.

1. Introduction

An ontology is a fundamental building component that is used to create more complex relationships between ideas. It lets members of a community of interest to create a shared language, which allows for more flexibility and aids in knowledge reusability and sharing.

What we are proposing is to recognize data units from our dataset such as names, including names of the individual, organization, and place, numerical terms including moment, date, occurrences such as battles, riots.

This paper presents valuable study for retrieving historical documents. Our goal is to create methods to automatically detect named entities.

2. Romanian Historical Data

There are digitized data for historical Romanian language. We had selected to examine a public historical document from the Bucharest Digital Library [1], ”534 Wallachia and Moldova Historical Documents Related to Transylvania.” We propose manual annotation data for tagging documents recording historical events, their participants, date, and place. [1]

3. Data field analysis and knowledge acquisition

Individuals belonging to a set of concepts and properties, for example, for instance - a specific tree is an individual for the concept of ”Tree”, while it can be stated that trees as a concept are material beings that have to be positioned on some location it is possible to state the specific location that a tree takes at some specific time [1][2].

Ontologies are often used to represent a specification of domain information by providing a consensual agreement on the semantic of the knowledge that the domain [4]. Another motivation behind using ontologies is that they allow for sharing and reuse of knowledge in a computational representation [2][7].

Queries within a document are usually limited to key-word searches. Relations between concepts within a document cannot be found by using a keyword search; we are only able to find the instances

Key words and phrases. historical ontologies, named-entity recognition, neural model.

[1] ©2019 Digital Library of Bucharest http://www.digibuc.ro/

Table 1. Named Entities in Historical Data

Entities	**Hierarchical Classes**		
Document	*Archive*	*Language*	*Signature*
Person	*Ancestor/Descendant*	*Parent/Child*	
Group	*Committee*	*Principality*	*Voivodship*
Place	*Location/Territory*		
Event	*Battle/Riot*	*Cultural/Social/Trade*	*Position/Rank*
Time	*Data*	*Period*	*Time-Interval*
	[1]		

[1] Digital Library of Bucharest

of the concepts contained in the document. For example, two instances, person X and person Y can be easily queried by keyword search; however, unless users read at least some parts of the document, they cannot determine whether these two people are related to each other, how this relationship is defined, and during what time period this relationship holds [4].

4. Results

In this paper, an attribute diagram as a working tool was applied for following informal model showed in Table 1.

The inference part, its correspondence, is like a database. Those that are deduced appear in yellow, but are deducted only when certain properties are defined. Based on the inferences, we still have some extra information that we don't have to enter.

On the other hand, in Protege, speaking of the fact that we only have binary relations, the attributes of an entity mean that they fulfill a bunch of binary relations, after which courts are placed there, unlike what we have in a database in which the structure is much simpler from this point of view. All those addresses in the IRI are for the unique identification of each entity.

4.1. Data annotation.

The dataset involves Romanian-language historical documents in text format [1]. Entities include AGENT, DOCUMENT, PERSON, LOCATION, EVENT, TIME and categories. This is an instance phrase that consists of each entity $[STARTtag] : [word] : [ENDtag]$.

This annotation method uses markup tags using angle brackets for defining named entities. Spaces are inserted before the phrase before $< START >$, those in the structure after the last word, after $< END >$; same as the punctuation marks.

Annotations are the transcribers remarks or clarifications to better clarify or comment on a portion of the transcription [7].

4.2. Data processing.

The majority of the time, people must categorize materials based on their context. It helps a lot when dealing with a huge quantity of documents. Therefore, it is very easy to use natural language processing (NLP) categorizer for the purpose we set out to do. There are several algorithms used in classifying these documents, most of which are semi-supervised algorithms such as maximum entropy, Naive Bayes [5][6].

Initially, we thought about binary classification, but the applications of binary classification are very limited, especially in the case of remote classification, where most classification problems involve more than two classes. Not having trained data, we tried to do it ourselves by manual

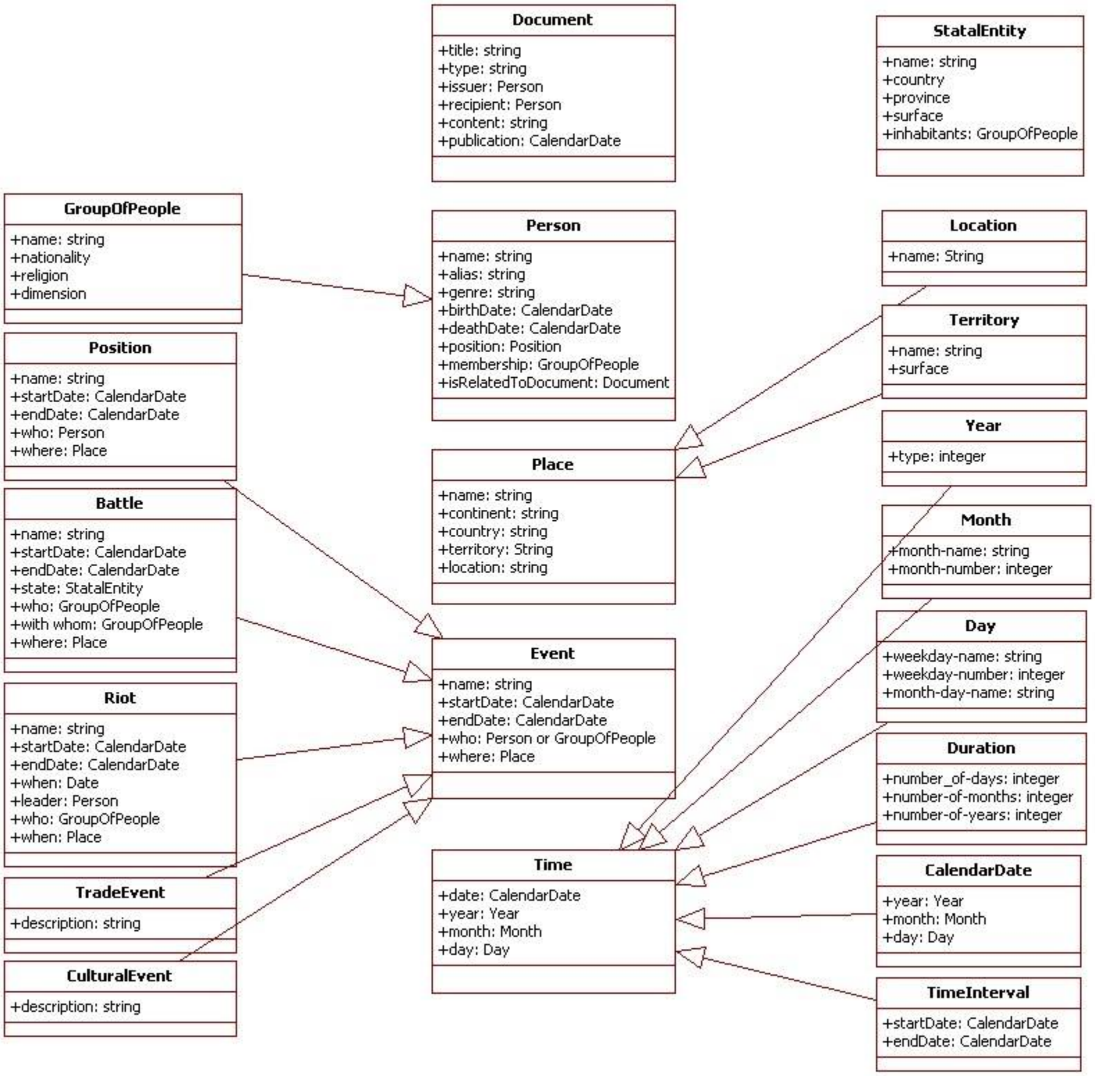

Figure 1. An extended view over the characteristics of the historical entities

annotation, and what is practiced is a training part and an evaluation part. By annotation, we processed the data in the file, 300 for training and 200 for evaluation.

We didn't focus on the functionality of the data, we focused on some data on which we have easy texts, even without diacritics, to see what is can get. Metainformations is more about storage, it is general in terms of content.

We train the entire network on one data set. If we give it words that are the Entities themselves, we take a class with a binary classification and see which entity we can get. We're taking unseen network input and that's it, we ask to be told what entity it is (extraction of the entity). We define a vector space in which each word has a position to vectorize our words. A neural network can only work with numbers, multiplies, fit forward, and we just have to interpret. To adjust, we go to a Gradient [3] (a convex function):

- 0 - set of matrices
- J - cost function

For Output, each neuron has a set of inputs (the first has only one, takes one feature and gives it to the other). If z is very large then -z is very small, and vice versa, resulting in the probability of belonging to a class (0,1) and with the minimization of the cost function.

The regularization constant C has several components:

(1) learning rate - tells us how far it takes us to Gradient [3]
(2) hiding the number - influences how much we process (auxiliary processing), each neuron can only make a binary classification; nonlinear hypotheses are denoted by h
(3) number of iterations: We have an array of words m, and we may not get a data set on a scroll, so we need to do more iterations. If we go through it several times and the function decreases cost does not help me to reach the output; set the parameters so that we make as few mistakes as possible, then we do a fit forward function for word training and data adjustment; if we don't have enough data it doesn't work properly, if we enter other inputs it will take time to get other outputs.

For the evaluation we divided the data that we annotated, the evaluation is done according to what we annotated manually on the second part, and then it is seen how good the results are, and if there is still the case for improvement it can be followed. The result is a comparison of the results.

5. Conclusion

In this study, a representation of historical concepts properties with relation classes and individuals was defined. We outlined a model for the annotation of digitized historical texts. This was possible independently of the platform, using annotated data text and integrating the resulting knowledge base file.

Our directions focus on processing historical languages by using both NLP and machine learning techniques. An analysis based on natural processing language would be needed to recognize and retrieves the name of historical entities. It is our intention, in the short term, to use those entities in order to build an ontology for historical data.

Acknowledgment

All authors are grateful to Professor dr. Parv Bazil, for the professional guidance given and valuable support.

References

[1] Cristea, D., Forascu, C., 2006, *Linguistic Resources and Technologies for Romanian Language*, Computer Science J. of Moldova, 14, 1 (40), pp.34–73.

[2] Katifori A., Kiyavitskaya N., Tympas A., 2009, *Ontologies for News and Historical Content*, Project Reference No. FP7–215874, Papyrus, (pp. 1–47).

[3] Ruder, S., 2016, *An overview of gradient descent optimization algorithms*, J. of arXiv:1609.04747v2.

[4] Yadav, P., 2017, *Semantic retrieval of historical documents based on IR approach*, International Journal of Applied and Natural Sciences (IJANS), 6(5), 51-58.

[5] Bird, S., Loper, E., Klein, E., 2009, *Natural Language Processing with Python*, O'Reilly Media Inc.

[6] Karttunen, M., Arnold, Zwicky, 1985, *Introduction to Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Cambridge University Press.

[7] Mirzaee Abar, V., 2004, *An Ontological Approach to Representing Historical Knowledge*, Vancouver, BC: University of British Columbia.

BABES-BOLYAI University of Cluj-Napoca, Faculty of Mathematics and Computer Science
Email address: danielacristea@cs.ubbcluj.ro, dana@cs.ubbcluj.ro, mihis.alin@cs.ubbcluj.ro