An Analysis of the Named Entity Recognition Problem in Digital Library Metadata

Nuno Freire[1,2], José Borbinha[1] and Pável Calado[1]
[1]IST/INESC-ID, Portugal
[2]The European Library, National Library of the Netherlands, Netherlands
{nuno.freire,jlb,pavel.calado}@ist.utl.pt

ABSTRACT

Information resources in digital libraries are usually described, along with their context, by structured data records, commonly referred as *metadata*. Those records often contain unstructured information in natural language text, since they typically follow a data model which defines generic semantics for its data elements, or includes data elements modeled to contain free text. The information contained in these data elements, although machine readable, resides in unstructured natural language texts that are difficult to process by computers. This paper addresses a particular task of information extraction, typically called named entity recognition, which deals with the references to entities made by names occurring in the texts. This paper presents the results of a study of how the named entity recognition problem manifests itself in digital library metadata. In particular, we present the main differences between performing named entity recognition in natural language and in the text within metadata. The paper finalizes with a novel approach for named entity recognition in metadata.

Categories and Subject Descriptors

E.1 [Data] Data Structures – Records; H.3.7. [Digital Libraries]; I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis; I.7.m [Document and Text Processing]:[Miscellaneous]

Keywords

Information extraction, entity recognition, metadata, digital libraries.

1. INTRODUCTION

A wide range of potentially usable business information exists in unstructured forms. Although this information is machine readable, it consists of natural language texts (it was estimated that 80% to 90% of business information may exist in those unstructured forms[1]).

As businesses become more data oriented, much interest has arisen in these unstructured sources of information. This interest gave origin to the research field of *information extraction*, which looks for automatic ways to create structured data from unstructured data sources [1]. An information extraction process can be characterized by an intention of selectively structure and combine data that is found in text, either explicitly stated or implied. The final output of the process will vary according to the purpose, but typically it consists in semantically richer data, which follows a structured data model, and on which more effective computation methods can be applied.

Information resources in digital libraries are usually described, along with their context, by structured data records. These data, which is commonly referred in the digital library community as *metadata*, may serve many purposes, and the most relevant being resource discovery. Those records often contain unstructured information in natural language text, which might be useful to judge about the relevance of the resource. The natural hypothesis is if that information can be represented with finer grained semantics, then the quality of the digital libraries is expected to improve.

This paper addresses a particular task of information extraction, typically called named entity recognition (NER), which deals with the textual references to entities, that is, when they are referred to by means of names occurring in natural language expressions, instead of structured data. This task deals with the particular problem of how to locate these references in the data set and how to classify them according their entity type [2].

In previous work, we have observed that current entity recognition and resolution techniques underperform in text within metadata [4][5]. In this paper we present a study on how named entities appear in metadata records, and on how structured metadata can be exploited to support the NER task, in order to improve the quality of the results.

The paper will follow with an introduction to NER and related work in Section 2. The data sets, which were the object of this study, are presented in Section 3, and the main results of the study are described in Section 4. The final section presents the conclusions of the study, proposes a novel approach for NER in metadata, and presents the future work.

2. PROBLEM AND RELATED WORK

The NER task refers to locating atomic elements in text and classifying them into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, etc. [2].

Initial approaches were based on manually constructed finite state patterns and/or collections of entity names [2]. However, named entity recognition soon was considered as a typical scenario for the application of machine learning algorithms, because of the potential availability of many types of evidence, which form the algorithm's input variables [3]. Current solutions can reach an F-measure accuracy around 90% in well-formed text, thus a near-human performance [2].

[1] http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551

JCDL'12, June 10–14, 2012, Washington, DC, USA.

However, previous work suggested that current NER techniques underperform when applied to unstructured texts existing within digital library metadata records [4][5]. Most research on NER has focused mainly on natural language processing, involving text tokenization, part-of-speech classification, word sequence analysis, etc. These techniques are, therefore, language specific and dependent of the lexical evidence given by the natural language text.

We are not aware of any studies specifically addressing the NER task in digital library metadata. The most similar scenarios, we are aware of, have researched information extraction techniques within relational database management systems, in order to improve the use of unstructured text inside relational databases [6]. This approach also was followed in [7], which addresses information extraction in a similar type of data with lack of lexical evidence, but applies simultaneously named entity recognition and entity resolution (the recognized names are resolved in a data set of known entities). The contribution of this work for advancing in NER techniques in this type of data is somewhat limited, since it was mainly focused on the recognition of entities that are present in the source data set.

3. STUDIED DATA SETS

Our study was performed on the data sets from Europeana[2], which consist in descriptions of digital resources of cultural interest. These data sets follow a data model using mainly Dublin Core elements, and named entities appear in data elements such as titles, textual descriptions, tables of contents, subjects, contributors and publication.

These data sets originate from several European data providers from the cultural sector, such as libraries, museums and archives. Several European languages are present, even within the description of the same resource. For example when the resource being described is of a different language than the one used to create its description. The data providers follow different practices for describing the digital resources. This environment makes the data sets to be highly heterogeneous in terms of data structure.[3]

We have studied two subsets of metadata from Europeana. The first subset is focused on the entity type location, which has been the focus of our previous work [5] on the recognition and resolution of this particular entity type. In this data set all references to locations have been manually annotated.

A second subset of the Europeana data sets has been manually annotated specifically for studying the general problem of NER in metadata. In this subset, we studied the three entity types on which most NER research has been focused, and which are commonly known as *enamex* [8]: person, location and organization.

4. STUDY RESULTS

Our main observation from the study of the data sets is that this kind of data typically presents different characteristics from well-formed text, which has been the focus of previous research in information extraction.

When applied to metadata without adaptations, the results of applying NER techniques designed for well-formed text will be negatively affected by three main limiting factors:

- The text within metadata may not contain enough lexical evidence to support the recognition of entities.
- The NER techniques designed for well-formed text can only be applied to metadata by first extracting the text, and then process it without taking in consideration the semantic context provided by the metadata record structure.
- If NER techniques use only the evidence available within the text, they will not take in consideration relevant evidence available in other structured data elements within the metadata record.

In the rest of this section we will present the results of our study that support our observations on each of these three limiting factors.

4.1 Availability of Lexical Evidence

In order to measure the extent to which the lack of lexical evidence influences the results of recognizing entities in digital libraries' metadata records, we have setup a small experiment on the recognition of the entity type *location*. We compared the results of two techniques that use lexical evidence for recognizing the location names, against a simple NER technique based on look ups of names in a geographical gazetteer.

The first technique tested was the implementation, provided by the OpenNLP package, of conditional maximum entropy models [9]. The second technique was the Stanford Named Entity Recognizer based on conditional random fields [10]. For both cases, we used the respective predictive models trained on the CoNLL 2003 training data [11]. The geographic gazetteer look up technique was developed by us for the purposes of this study. It was designed to perform a simple matching of the location names in the texts without using any lexical evidence, so that we could analyse to what extent the lack of lexical evidence affects the quality of the NER results.

The experiment was conducted by processing, with the three techniques, a collection of records from the Europeana data sets. The techniques were evaluated according to the *exact-match* method, which has been used in several named entity recognition evaluation tasks [11]. In the *exact-match* method, a named entity is only considered correctly recognized if it is exactly located as in the manual annotation. Recognition of only part of the name, or with words that are not part of the name, is not considered correct. In combination with the *exact-match* method, we used the metrics of precision[4], recall[5] and F_1-measure[6].

This collection is a subset of the evaluation collection described in [5], which contains 752 metadata records with a total of 2823 annotated location names. For this study, we only used records in the languages supported by the NER techniques we tested. Therefore, it was restricted to 372 records in English, Spanish, Dutch and German.

The results are presented in Table 1. Both techniques that used lexical evidence achieved higher precision than the gazetteer look up technique, but performed much lower in terms of recall, resulting in considerably lower F_1-measure, with less 0.20 for the Maximum Entropy ($P<0.001$), and less 0.12 for the CRF ($P<0.01$).

[2] http://www.europeana.eu/

[3] http://dublincore.org/

[4] The percentage of correctly identified named entities in all named entities found

[5] The percentage of named entities found compared to all existing named entities

[6] The weighted harmonic mean of precision and recall (equal weights for recall and precision)

Table 1. Location recognition results on a collection of 372 metadata records

Technique	Precision	Recall	F1-measure
Conditional Maximum Entropy	0.95	0.29	0.45
Conditional Random Fields	0.95	0.38	0.53
Geographic Gazetteer Lookup	0.92	0.53	0.65

These results support our hypothesis that the lack of lexical evidence in the text of metadata elements negatively affects the results of NER techniques. It can also be observed that it mainly affects recall.

4.2 Semantic Context of the Metadata Elements

Each metadata element provides its own semantic context to the text contained within, and this context influences the types of entities that are referenced.

To investigate this aspect we have analysed the occurrence of referenced entity types found in the Europeana data sets. These data sets contain several different data elements with different associated semantics, allowing us to study the extent to which the semantic context influences the entity types of named entities. In this study we addressed the three entity types on which most NER research has been focused: person, location and organization.

This collection was created by randomly selecting records in the English language. The selection process was done in two steps: first, all records in the English language were selected from all Europeana data providers; and second, a random selection of records was performed, balancing the number of records chosen across different providers.

The evaluation data set was manually annotated. In very few cases, the manual annotation was uncertain, because the metadata records may not contain enough information to support a correct annotation. For example, some sentences with named entities were too small and no other information was available in the record to support a decision on the annotation of the named entities to their entity type. Annotation was performed as follows:

- Named entities were annotated with their entity type: *person*, *location* or *organization*;
- If the annotator was unsure of the entity type of a named entity, he would annotate it as *unknown*. These annotations were not considered for our study.

In total, the evaluation data set consisted in 120 records containing in its elements 584 references to persons, 457 to locations and 153 to organizations.

For each element, the distribution of references to different entity types was calculated, and the results can be visualized in Table 2. We can observe that the elements *title* and *description* have a similar distribution of entity types, and notice that persons were the most frequently found entity type. The element *subject* and *table of contents* also share a similar distribution, but with a higher frequency for locations. All other elements presented distinct distributions. In the *creator* element, only persons and organizations were found, in the element *coverage* exclusively locations were found, and in element *publisher* mostly organizations and locations were found. Such difference in distributions of entity types, if made available as evidence to the NER algorithm, can support the recognition of a name, and the disambiguation of the entity type. For example, if the name *Washington* is found, although it can be of type *person* of *place*, if found in a *coverage* element it likely refers to a place, while if found in a *creator* element, it likely refers to a *person*.

Table 2. Distribution of textual references to entities by entity type across data elements of the Europeana data set

Element	Locations	Persons	Organiz.	Total
Title	86 (34%)	142 (56%)	26 (10%)	254
Creator/Contributor	0 (0%)	156 (85%)	27 (15%)	183
Subject	136 (64%)	60 (28%)	16 (8%)	212
Coverage	79 (100%)	0 (0%)	0 (0%)	79
Publisher	52 (44%)	17 (15%)	48 (41%)	117
Table of Contents	29 (69%)	10 (24%)	3 (7%)	42
Description	75 (24%)	199 (65%)	33 (11%)	307
Total	**457**	**584**	**153**	**1194**

4.3 Evidence from Structured Data within the Metadata Records

Application of the NER techniques designed for well-formed text in metadata, does not take advantage of the semantic context provided by the metadata record structure.

Metadata records however, have available not only the text of the data elements, but also other data elements which may be structured. These structured elements may be exploited during the NER process, allowing the use of semantic information to support a correct recognition of the named entities.

An example of one such case was observed in our study of the named entities of the entity type *location* within the Europeana data sets (in the data sets described in Section 4.1). In these data sets, a structured data element contained structured data about the provenance of the metadata: the country code of the institution from where the data record originated.

In order to study the possible influence of making the country of provenance available to the NER algorithms, we have analysed the relationship between the country of provenance of the data and the correct resolution for the names of locations. In this data set, the named locations are annotated also with their correct resolution in a geographic gazetteer, allowing us know the country corresponding to each named location.

In the data sets, we encountered 54% of the references that matched the country of origin of the data, 9% matched a neighbour country, and 37% were from the rest of the world. We tested the hypothesis that the probability of a named location to be from the country of provenance of the data, is higher that the probability of being located in the rest of the world. The differences in the percentages were found to be statistically significant, with $P<0.01$.

Several different metadata models are used in digital libraries, therefore the availability of structured data for use in NER will vary in each case. But these results support that the structured data may be used to improve the NER results.

Table 3. Relationship statistics between named locations and the country of provenance of the metadata in the Europeana data set

Total named locations	In the same country	In a neighbour country	Rest of the World
2823	1524 (54%)	255 (9%)	1044 (37%)

5. CONCLUSIONS AND FUTURE WORK

Based on the results of the study presented in the previous Section, we believe that an approach for NER on digital libraries' metadata should follow the following general guidelines:

- The data record should be available throughout the entire NER process. If the processing exploits the data record, instead of just the text within, more evidence will be available, allowing better recognition of the named entities.
- The approach must capture lexical and non-lexical evidence from the text. This would allow process to handle cases of data containing well-formed text, as well as simple textual expressions.
- The approach must be adaptable. Different data sets will present different textual patterns and languages, therefore the evidence to support the recognition and resolution of entities will vary considerably in different data sets. Supervised machine learning based approaches are the state of the art in both entity recognition and resolution in well formed, and the adaptability of these algorithms is just as relevant for the large variety of metadata models in use.

Based on this analysis, we believe that improvements for NER in metadata can be achieved over using NER algorithms for well-formed text. The results of this study provide evidence which motivate the investigation of adaptations to the state of the art NER techniques, and in the methods of their application in metadata.

Our proposal for a general NER process for metadata is depicted in **Error! Reference source not found.**. The first design decision of this process is to start by having as input a metadata record, instead of a data element or the text previously extracted from a data element or record. This allows the process to have the structured record available throughout the entire process, for providing evidence.

The first task of the process consists in selecting the target data elements where entity names are to be recognized and resolved. It is then followed by a task of basic text processing, in order to transform the text within the data elements into a sequence of tokens (words, punctuation marks, numbers, etc.).

In the third task, the text tokens are analyzed in order to extract any possible evidence that may support the recognition of the entities names. Typically, this task includes lexical analysis of sentences (e.g. part-of-speech tagging), analysis of words features (e.g. capitalization), checking the existence of the tokens in collections of entity names, textual patterns detection, etc. The data element itself may provide evidence as well, since it provides a semantic context for the interpretation of the text. Also extra evidence can be extracted from other structured elements on the same record.

The following, and final task, is the recognition of names of entities. It consists in reasoning on the sequence of tokens and the associated evidence gathered in the previous task. Its outcome consists in the recognized names and their respective positions in the source text. We advocate that this task needs to be performed by machine learning based techniques, in order to have an approach that is adaptable to different data sets and data elements with different evidences and textual patterns.

In future work we will conduct experiments with implementations of this proposed NER process. These experiments will be designed to allow the analysis of the performance of state of the art techniques in metadata, and evaluate the use of alternative methods that may provide improvements.

6. REFERENCES

[1] S. Sarawagi, “Information Extraction”, Found. Trends databases, vol. 1, pp. 261-377, Now Publishers Inc., 2008.

[2] D. Nadeau, S. Sekine, “A survey of named entity recognition and classification”, Linguisticae Investigationes, volume. 30, number 1, pp. 3-26, John Benjamins Publishing Company, 2007.

[3] A. McCallum, D. Freitag, and F. C. N. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation”, 17th International Conference on Machine Learning, pp. 591-598, 2000.

[4] B. Martins, J. Borbinha, G. Pedrosa, J. Gil, N. Freire, “Geographically-aware information retrieval for collections of digitized historical maps”. 4th ACM Workshop on Geographical information Retrieval, 2007.

[5] N. Freire, J. Borbinha, P. Calado, B. Martins, "A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records", ACM/IEEE Joint Conference on Digital Libraries, 2011.

[6] P. King, A. Poulovassilis, “Enhancing database technology to better manage and exploit Partially Structured Data”. Research Report pjhk/ap,11/2000, University of London, 2000. Available at http://www.dcs.bbk.ac.uk/research/techreps/2000/bbkcs-00-14.pdf.

[7] D. Williams: Combining Data Integration and Information Extraction. PhD thesis, University of London (2008)

[8] R. Grishman, B. Sundheim, "Message Understanding Conference - 6: A Brief History", International Conference on Computational Linguistics, 1996.

[9] J. Goodman, “Sequential Conditional Generalized Iterative Scaling”, 40th Annual Meeting of the Association for Computational Linguistics, pp. 9-16, 2002.

[10] J. R. Finkel, T. Grenager, C. Manning, “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370, 2005.

[11] E. Sang, F. Meulder, “Introduction to the CoNLL-2003 shared task: language-independent named entity recognition,” seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, pp. 142-147, 2003.

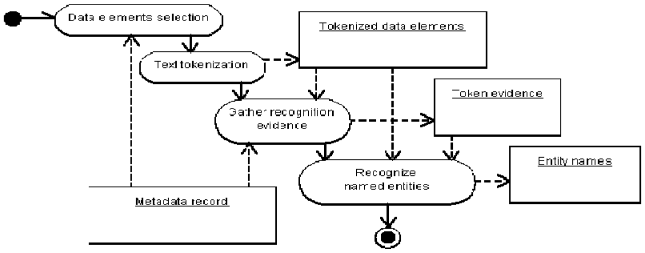

Figure 1. A general process for named entity recognition in metadata