An Approach for Named Entity Recognition in Poorly Structured Data

Nuno Freire[1,2], José Borbinha[1], and Pável Calado[1]

[1] INESC-ID/Instituto Superior Técnico - Technical University of Lisbon,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2] The European Library, National Library of the Netherlands,
Willem-Alexanderhof 5, 2509 LK The Hague, Netherlands
{nuno.freire,jlb,pavel.calado}@ist.utl.pt

Abstract. This paper describes an approach for the task of named entity recognition in structured data containing free text as the values of its elements. We studied the recognition of the entity types of *person*, *location* and *organization* in bibliographic data sets from a concrete wide digital library initiative. Our approach is based on conditional random fields models, using features designed to perform named entity recognition in the absence of strong lexical evidence, and exploiting the semantic context given by the data structure. The evaluation results support that, with the specialized features, named entity recognition can be done in free text within structured data with an acceptable accuracy. Our approach was able to achieve a maximum precision of 0.91 at 0.55 recall and a maximum recall of 0.82 at 0.77 precision. The achieved results were always higher than those obtained with Stanford Named Entity Recognizer, which was developed for grammatically well-formed text. We believe this level of quality in named entity recognition allows the use of this approach to support a wide range of information extraction applications in structured data.

Keywords: named entity recognition, structured data, metadata, conditional random fields.

1 Introduction

A wide range of potentially usable business information exists in unstructured forms. Although that 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] [2]).

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 [3]. 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

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 718–732, 2012.

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 data 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 system 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 [4].

We describe a NER approach, which we studied on the particular case of metadata from the cultural heritage domain, represented in the generic Dublin Core[1] data model, which typically contains uncontrolled free text in the values of its data elements. We refer to this kind of data as poorly structured data. Typical examples of such data elements are the titles, subjects, and publishing information.

NER has been extensively researched in grammatically well-formed text. In poorly structured data however, the text may not be grammatically well-formed, so our assumption is also that the data structure provides a semantic context which may support the NER task.

This paper presents an analysis of the NER problem poorly structured data, describes a novel NER approach to address this kind of data, and presents an evaluation of the approach on a real set of data. The paper will follow with an introduction to NER and related work in Section 2. The proposed approach is presented in Section 3, and the evaluation procedure and results are presented in Section 4. Section 5 concludes and presents 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. [4].

Initial approaches were based on manually constructed finite state patterns and/or collections of entity names [4]. 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 [5]. Current solutions can reach an F-measure accuracy around 90% [4] in grammatically well-formed text, thus a near-human performance.

[1] http://dublincore.org/

However, previous work suggested that current NER techniques underperform when applied to texts existing within structured digital library records [6] [7] [8]. Most research on NER has focused mainly on natural language processing, involving text tokenization, part-of-speech classification, word sequence analysis, etc. Recognition with these techniques is therefore language specific and dependent of the lexical evidence given by the natural language text.

The most similar scenarios we are aware of have researched information extraction within poorly structured data, and with a different focus than us. Research described in [9] proposes the use of information extraction techniques within relational database management systems, in order to exploit existing unstructured data within databases. This approach also was followed in [10], which addresses information extraction in a similar type of data as we do, 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 only addressed the recognition of entities that are present in the source data set. Similarly to our experience, this work also reports difficulties with the NER solutions for natural language text (although only one tool was evaluated [11]). However, this approach differs significantly from ours. In order to improve the NER results, this approach was based on the evidence provided by structured data about the entities to be recognized, and the recognition model is based on manually crafted parsing rules created by a domain expert.

Although not addressing the same type of data as we do, we can find approaches used in other contexts that also perform NER in text containing little or no lexical evidence. In [12], an approach is described for performing information extraction on a particular kind of unstructured and ungrammatical text posted on the World Wide Web, such as item auction posts or online classifieds. The aim of this approach however is to extract a structured data record from each post, assuming that each post contains multiple attributes' values of one entity, making the approach not applicable to our scenario.

Other works, addressing NER in text without lexical evidence, focused on search engine queries [13][14]. In this work the problem is defined assuming the existence of one main entity per query, and adopted a specific technique for such cases, based on query logs [13] or user sessions [14] and topic models. We find the topic model approach to be not generally applicable for NER in to the data we are studying, since it assumes the existence of only one main entity per data element value.

3 Approach

We aimed at developing a general NER approach which could be systematically applied to any poorly structured data set. This section starts by presenting our analysis of the NER problem in structured data, and the general design decisions behind our approach. The description of the approach follows, and finalizes with the description of relevant implementation details.

3.1 Analysis

From our analysis of named entities found in structured data sets, we can highlight the following points:

- Availability of lexical evidence varies in many cases. In some data elements we found grammatically well-structured text, in other elements we found short sentences, containing very limited lexical evidence, or plain expression with practically non-existing lexical evidence. We also observed that in some cases, analysis of the same field across several records, revealed a mix of all cases.
- Instead of lexical evidence, we observed that, in some cases, textual patterns are often available and could be explored as evidence for NER. For example, punctuation marks play an important role, but its use may differ from how they are used in natural language text.
- These data elements are typically modeled with general semantics. The semantics associated with each element influences the type of named entities found in the actual records. Therefore, we observed different probability distributions for each entity type across data elements.
- One of the major sources of evidence is the actual name of the entities. Each entity type presents names with different words and lengths, and also with different degrees of ambiguity with other words and entity types.

From this analysis we believe that a generic approach must be highly adaptable, not only to the data set under consideration but also to each data element. Text found in each element across the whole data set is likely to be associated with particular patterns and degrees of available lexical evidence.

On a more generic level, the approach should have a strong focus on the disambiguation of the names between the supported entity types, and be able to disambiguate between entity names and other nouns/words.

3.2 Entity Types

We studied the three entity types on which most NER research has been focused, and which are commonly known as *enamex* [15]: person, location and organization. In addressing these three entity types, we wanted to design an approach that was not limited to a set of known entity names, but could recognize any named entity of the supported entity types, as usually done in NER in grammatically well-structured text.

As mentioned in the previous section, in structured data the characteristics of the names of persons, organizations, and places are a strong evidence for recognizing the named entities and determining their entity type. Therefore, in order to allow the predictive model to use the likelihood of a token being part of a named entity, we have collected name usage statistics from comprehensive data sets of persons, organizations and locations.

Person and organization name statistics were extracted from VIAF - Virtual International Authority File [16]. VIAF is a joint effort of several national libraries from all continents towards a consolidated data set gathered for many years about the creators of the bibliographic resources held at these libraries.

Location name statistics were extracted from Geonames [17], a geographic ontology that covers all countries and contains over eight million locations.

A description of how the statistics were extracted, and used in the predictive model, is presented in Section 3.4.

3.3 Predictive Model

Our analysis suggested that a flexible approach with the capacity to adapt to the data set would be necessary for performing NER in structured data. This suggested the application of a machine learned model, an option also supported by the literature review of state of the art NER approaches.

The NER problem can be formulated as follows. Given a text string x and a set of entity types Y, where x consists of a sequence of tokens $x_1 \ldots x_n$, and each token is a word or a punctuation mark, the entity recognition task consists in segmenting x into a sequence s of non-overlapping segments $s_1 \ldots s_p$ where each segment s_j is associated with a $y_j \in Y$, and a start position t_j , and an end position u_j (for notation readability purposes we assume Y to also contain a *non_entity* type). All segments of s are non-overlapping and fully encompass all tokens of x, therefore for all x_i exists one and only one s_j that satisfies $s_{tj}<= i$ *and* $s_{uj}>=i$.

We use as a basis the conditional models of conditional random fields (CRF) [18]. CRFs define a conditional probability $p(y|x)$ over label sequences given a particular observation sequence x. These models allow the labelling of an arbitrary sequence x' by choosing the label sequence y' that maximizes the conditional probability $p(y'|x')$. The conditional nature of these models allows arbitrary characteristics of the sequences to be captured by the model, without requiring previous knowledge, by the modeller, about how these characteristics are related [19].

In order to find the sequence s that correctly recognizes the entity names from the observation sequence x, evidence is extracted or calculated. This evidence consists in a set of features which capture those characteristics of the empirical distribution of the data that support the recognition of names. Many different methods have been used to calculate and use features in a combined manner. Features may be calculated from natural language processing of the source text, by rules defined by domain experts, by lookups in lists of entity names and ontologies, from syntactical characteristics of the tokens, etc. The following section presents the set of features that we defined for our particular predictive model.

3.4 Features

Several features were defined to give the predictive model the capability to capture distinct aspects of the text, such as locating potential names, disambiguate between entity types and other words, or detecting textual patterns from syntactical and lexical evidence. This section presents the definition of these features.

A set of features were defined to provide the predictive model with some evidence for locating potential names of entities in the text. These features were created based on data or statistics taken from the comprehensive listings of names described in Section 3.2. Each entity type has different characteristics in the way entities are named, so we defined the features in different ways for each entity type.

The features for person names explore how frequent a word was found in person names, making a distinction between first names, surnames and names that appear in lowercase. Let *F* denote a bag built from all first names found in VIAF, and let *S* denote a bag built from all surnames found in VIAF, and let *C* be a bag built from all names found non-capitalized in VIAF. We define the following real valued features:

$$personFirstName(x, i) = \log\left(1 + {}^{F_{\#x_i}} \Big/ \left(\frac{\sum_{j=0}^{\#F} F_{\#j}}{\#F}\right)\right)$$

$$personSurname(x, i) = \log\left(1 + {}^{S_{\#x_i}} \Big/ \left(\frac{\sum_{j=0}^{\#S} S_{\#j}}{\#S}\right)\right)$$

$$personNoCapitalsName(x, i) = \log\left(1 + {}^{C_{\#x_i}} \Big/ \left(\frac{\sum_{j=0}^{\#C} C_{\#j}}{\#C}\right)\right)$$

For organizations, only one feature was defined. Let *C* be a bag built from all words and punctuation marks found in the names of organizations in VIAF, we define the following real valued feature:

$$organizationName(x, i) = \log\left(1 + {}^{C_{\#x_i}} \Big/ \left(\frac{\sum_{j=0}^{\#C} C_{\#j}}{\#C}\right)\right)$$

For places, the diversity of the names makes the frequency of use of the words not effective, so one feature was defined, using the type of geographic entity and the highest population known for a place on whose name the word appears in. Let *C* denote a bag built from all tokens found in the names of continents and countries. Similarly let *D, E, F* and *G* denote bags built from all tokens found in the names of cities, administrative divisions or islands, natural geographic entities, and other geographic features, respectively. Also let $population(t) \longmapsto \mathbb{N}$ denote a function that returns the maximum population found in a location name with token *t*. We defined the following real valued feature:

$$locationName(x,i) = \begin{cases} 1, if\ x_i \in C \\ \dfrac{\min(100000,\ \text{population}(x_i))}{100000}, if\ x_i \in D \\ 0.7, if\ x_i \in E \\ 0.6, if\ x_i \in F \\ 0.1, if\ x_i \in G \\ 0, otherwise \end{cases}$$

Some features are based on data extracted from the WordNet [20] of the language matching the language of the source text, which in the case we studied was English. These features provide evidence to disambiguate between named entities of the target types and other words.

With the aim to disambiguate between proper nouns referring to other entity types, and proper nouns referring to persons, locations and organizations, we define the feature $properNoun(x,i) \longmapsto \{0,1\}$. Let *P* denote the set of all variants in synsets which have a part-of-speech value of *proper noun*, and let *G, H, I, J, K, L* denote the sets of variants in synsets which are hyponyms, either directly or transitively, of one of the synsets[2] *geographic area#noun#1, landmass#noun#1, district#noun#1, body of water#noun#1, organization#noun#5, and person#noun#1*, respectively. The feature is defined as:

$$properNoun(x,i) = \begin{cases} 1, if\ x_i \in P \backslash (G \cup H \cup I \cup J \cup K \cup L) \\ 0, otherwise \end{cases}$$

We also use the Wordnet to capture the possible part-of-speech of some tokens. We defined the feature $posNoun(x,i) \longmapsto \{0,1\}$, which indicates if the token exists in a synset with part-of-speech *noun*. Let *A* denote the set of variants in synsets which have a part-of-speech value of *noun,* we define the feature as:

$$posNoun(x,i) = \begin{cases} 1, if\ x_i \in A \\ 0, otherwise \end{cases}$$

Similar features were defined for other parts-of-speech: $posVerb(x,i) \longmapsto \{0,1\}$, $posAdjective(x,i) \longmapsto \{0,1\}$, $posAdverb(x,i) \longmapsto \{0,1\}$, and $posPreposition(x,i) \longmapsto \{0,1\}$.

We also defined features to capture syntactical characteristics of the text and the tokens. The features $startOfElement(x,i) \longmapsto \{0,1\}$ and $endOfElement(x,i) \longmapsto \{0,1\}$ indicate if token x_i is at the start or at the end of the value of the data element. The case of the token is captured through the features $isCapitalized(x,i) \longmapsto \{0,1\}$ and $isFullCaps(x,i) \longmapsto \{0,1\}$, which indicate if the token is a word and contains the first letter in uppercase, or all letters in uppercase, respectively. The token's character length is captured by the feature $tokenLength(x,i) \longmapsto \mathbb{N}$.

The tokens are also used in a nominal feature $token(x,i) \longmapsto T$, where *T* denotes the set of tokens built from the three preceding tokens, and the two following tokens, of every named entity found in the training data:

[2] To refer to Princeton WordNet synsets, we use the notation w#p#i where *i* corresponds to the *i*-th sense of a literal *w* with part of speech *p*.

$$token(x,i) = \begin{cases} x_i, if\ x_i\ \in T \\ \emptyset, otherwise \end{cases}$$

Capitalization statistics of words in the data set are extracted and used in a feature. Let C denote the bag of capitalized words in the data set, and let D denote a bag of the non-capitalized words in data set, we define the following real valued feature:

$$capitalizedFrequency(x,i) = \log\left(\frac{C_{\#x_i}}{(1 + D_{\#x_i})}\right)$$

Since typically each data element will have values with different characteristics, a feature is necessary to capture the data element where the text is contained. We defined the feature $dataElement(x,i) \longmapsto D$, where D denotes the set of data element identifiers of the data model (for example, in data encoded in XML, these identifiers consist of the xml element's namespace and element's name).

Additional features are defined in similar way, but they refer to the three previous tokens and the two following tokens, instead of the current one.

3.5 Implementation Details

In this section we provide some relevant details of the implementation of our approach, in particular we address text tokenization and the CRF implementation and configuration.

Tokenization of the text inside the data elements is performed only at word level. No sentence or paragraph tokenization is performed, since in many cases well-structured sentences are not present in the data and the results of sentence and paragraph tokenization could invalidate the detection of patterns in the data.

Word tokenization is performed in a language independent way. We also justify this option to avoid the breaking of patterns in the data, in particular in cases where punctuation is used in the data with different meanings than it is has in natural language text. We have applied the word breaking rules of UNICODE [22].

The CRF implementation used was provided by the Java implementation in the MALLET - Machine Learning for Language Toolkit [21]. The CRF was configured to use the three previous states in the sequence in the labelling of the sequence, and was trained using an objective function for CRFs that consists in the label likelihood plus a Gaussian prior on parameters.

4 Evaluation

The evaluation of our approach was performed in the data sets from Europeana[3], which consist in descriptions of digital objects of cultural interest. This data set follows a data model using mainly Dublin Core elements, and named entities appear in

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

data elements for titles, textual descriptions, tables of contents, subjects, authors and publication.

The data set contains records originating from several European providers from the cultural sector, such as libraries, museums and archives. Several European languages are present, even within the description of the same object, for example when the object being described is of a different language than the one used to create its description.

Providers from where this data originates follow different practices for describing the digital objects, which causes the existence of highly heterogeneous data. Lexical evidence is very limited in this data set, so it provides a good scenario for the evaluation of the evidence made available by the structure and textual patterns of the data.

This section describes the experimental setup and its results. It will follow with the description of the data set used for evaluation, and then describe the evaluation procedure. Results of the evaluation are presented afterwards, and it finalizes with the results of the evaluation of individual features.

4.1 Evaluation Data Set

An evaluation of our approach was performed on a selected collection of metadata records from Europeana. 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.

In total, the evaluation data set[4] consisted in 120 records containing in its elements 584 references to persons, 457 to locations and 153 to organizations, as shown in Table 1.

Table 1. Data elements studied in the data set and total annotated named entities

Data element	Element definition[5]	Pers.	Locat.	Organiz.
Title	A name given to the resource.	142	86	26
creator / contributor	An entity primarily responsible for making the resource / An entity responsible for making contributions to the resource.	156	0	27
Subject	The topic of the resource.	60	136	16
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.	0	79	0
Description	An account of the resource.	199	75	33
table of contents	A list of subunits of the resource.	10	29	3
Publisher	An entity responsible for making the resource available.	17	52	48
	Total:	**584**	**457**	**153**

[4] The data set is available for research use at http://web.ist.utl.pt/~nuno.freire/ner/

[5] Element definitions were taken from the Dublin Core Metadata Terms.

The evaluation data set was manually annotated. In very few cases, the manual annotation was uncertain, because the data 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 classification of the named entities to their entity type. Named entities were annotated with their *enamex* type. If the annotator was unsure of the *enamex* type of a named entity, he would annotate it as *unknown*. These annotations were not considered for the evaluation of the results, and any recognition made in these entities was discarded.

4.2 Evaluation Procedure

The accuracy of the results of our approach was compared with that of other two approaches: one was the implementation of a conditional maximum entropy model [25], taken from the OpenNLP package; the other was based on conditional random fields [26], from the Stanford Named Entity Recognizer (Stanford NER). For both cases, we used the respective predictive models trained on the CoNLL 2003 English training data [27]. However, since in all tests the Stanford NER performed better than OpenNlp, for readability, we only present the results of Stanford NER as our baseline for comparison.

Since our predictive model was trained on the evaluation data set, all the measurements were obtained using cross-validation tests, which has been widely accepted as a reliable method for calculating generalization accuracy [24]. Cross-validation involves partitioning the evaluation data set into complementary subsets, testing on one subset, while training on the remaining subset. Ten-fold cross-validation was performed using different partitions, and the validation results were averaged over the ten runs.

As the NER evaluation method, we have used the *exact-match* method. This method has been used in several named entity recognition evaluation tasks [23] [27]. In the *exact-match* method, an 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[6], recall[7] and F_1-measure[8].

To evaluate on the balance between results in precision and recall, we have taken measures at several minimum confidence thresholds. For both our approach and the baseline, we only consider a named entity recognized if the joint probability of the corresponding segment is equal or above the minimum confidence threshold.

4.3 Results

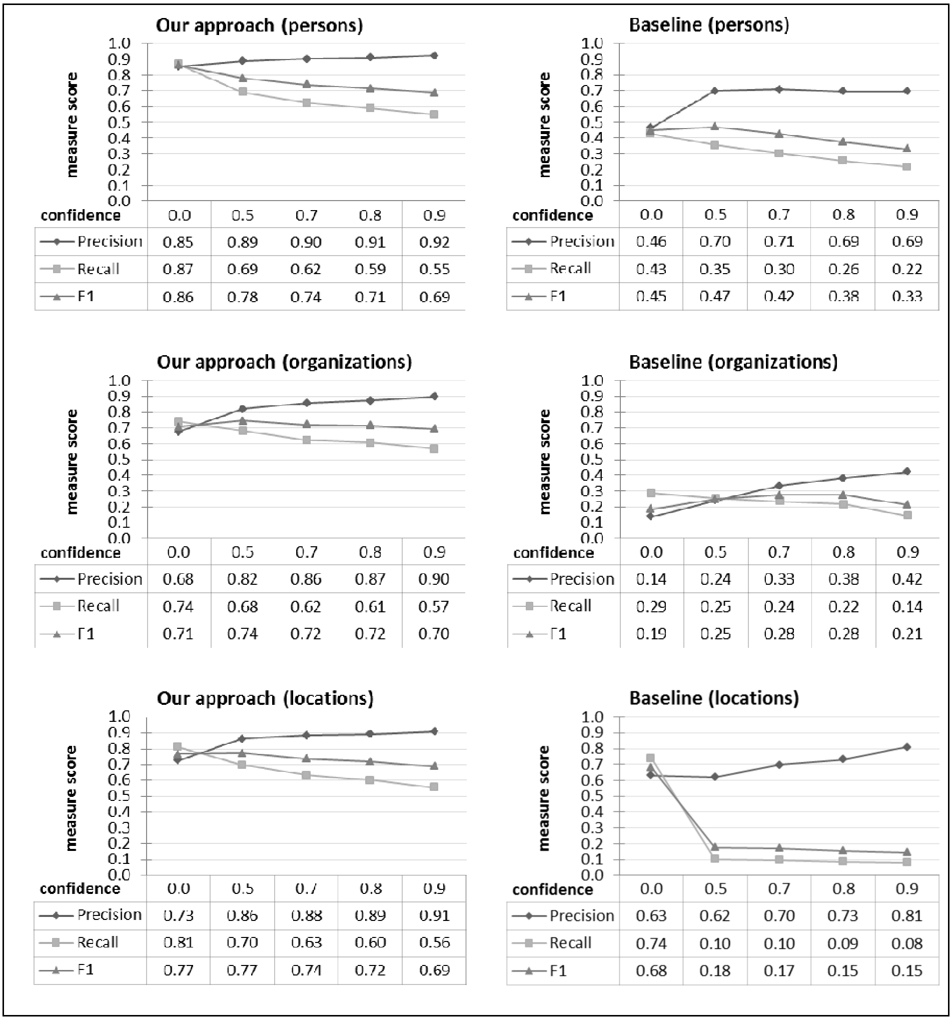
The overall results of the evaluation of all entity types are presented in Fig. 2, and the results of each entity type are presented in Fig. 1. The results of our approach were

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

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

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

higher for all entity types, metrics and confidence levels. The differences between our approach and the baseline were statistically significant with P>0.001 for all measurements except for the entity type location where, in the lowest confidence threshold, we obtained P>0.01 on the three metrics.



Our approach (persons)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.85	0.89	0.90	0.91	0.92
Recall	0.87	0.69	0.62	0.59	0.55
F1	0.86	0.78	0.74	0.71	0.69

Baseline (persons)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.46	0.70	0.71	0.69	0.69
Recall	0.43	0.35	0.30	0.26	0.22
F1	0.45	0.47	0.42	0.38	0.33

Our approach (organizations)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.68	0.82	0.86	0.87	0.90
Recall	0.74	0.68	0.62	0.61	0.57
F1	0.71	0.74	0.72	0.72	0.70

Baseline (organizations)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.14	0.24	0.33	0.38	0.42
Recall	0.29	0.25	0.24	0.22	0.14
F1	0.19	0.25	0.28	0.28	0.21

Our approach (locations)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.73	0.86	0.88	0.89	0.91
Recall	0.81	0.70	0.63	0.60	0.56
F1	0.77	0.77	0.74	0.72	0.69

Baseline (locations)

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.63	0.62	0.70	0.73	0.81
Recall	0.74	0.10	0.10	0.09	0.08
F1	0.68	0.18	0.17	0.15	0.15

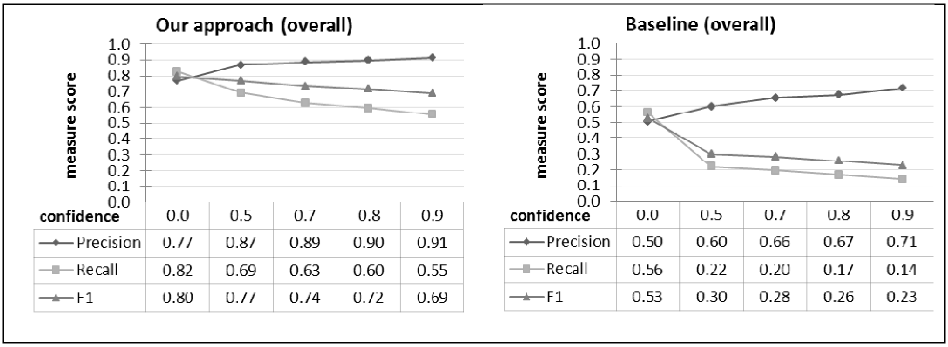
Fig. 1. Precision, recall and F_1 results of the three *enamex* entity types measured on the evaluation data set

Both Stanford NER and our approach are based on CRFs. Although the implementation of CRFs used was not the same, and other differences exist on how CRFs are used, we believe that the difference in the results obtained with both approaches is due to the different features used, therefore supporting our initial hypothesis that the semantic context of the data structure, and non-lexical features, could support NER.

An interesting result can be observed at the lowest confidence threshold result for the entity location, where Stanford NER was actually able to achieve a F_1 of 0.68, but the probability given by the CRF predictive model was close to zero for more than 70% of the recognized named entities. This observation suggests that the lack of lexical evidence had a major impact in its results.

Results of both approaches generally showed lowest values for recall than for precision. In our approach overall recall ranged from 0.55 to 0.82 while overall precision ranged from 0.77 to 0.91. Given the importance of the features based on the names of entities, as show in the next section, we believe that the lower recall is mainly caused by names that had no presence in the entity names data sets. However, we were not able to empirically support this conclusion.

Our approach was able to achieve a high precision of 0.91 at 0.55 recall, or reach a recall of 0.82 at 0.77 precision. We believe these values reached levels high enough to support a wide range of information extraction applications, which may have different requirements for recall or precision.



confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.77	0.87	0.89	0.90	0.91
Recall	0.82	0.69	0.63	0.60	0.55
F1	0.80	0.77	0.74	0.72	0.69

confidence	0.0	0.5	0.7	0.8	0.9
Precision	0.50	0.60	0.66	0.67	0.71
Recall	0.56	0.22	0.20	0.17	0.14
F1	0.53	0.30	0.28	0.26	0.23

Fig. 2. Precision, recall and F_1 results of all entity types measured on the evaluation data set

4.4 Feature Evaluation

In order to evaluate the contribution of each feature for the quality of the NER results, we have performed a feature selection evaluation based on the wrapper methodology [28]. This method employs cross-validation using the actual target learning algorithm to estimate the accuracy of subsets of features.

To conduct this evaluation, we have grouped related features (for example, all features related with part-of-speech were considered one group), and performed an exhaustive evaluation for all combinations of groups of features.

In total, we formed 10 groups of features and tested all combinations of 7 groups. Each feature combination was evaluated by a 10-fold cross-validation test and the best performing feature combination, measured by the F_1, of each fold was noted.

Table 2 summarizes the results, by showing how often each feature was present in the best performing combination of the 10 folds, for all entity types, and for each type individually. Since the features related with the names of the entities were essential for the overall results, on the evaluation on the individual entity types, we always used combinations including these three groups of features, so that the results could be more easily compared and analyzed.

Table 2. Results of the evaluation of the features

Feature groups	Included in best combination			
	all types	persons	locations	organizations
$\boldsymbol{personFirstName(x,i)}$ $\boldsymbol{personSurname(x,i)}$ $\boldsymbol{personNoCapitalsName(x,i)}$	100%	100%	100%	100%
$\boldsymbol{organizationName(x,i)}$	100%	100%	100%	100%
$\boldsymbol{locationName(x,i)}$	100%	100%	100%	100%
$\boldsymbol{token(x,i)}$ $\boldsymbol{startOfElement(x,i)}$ $\boldsymbol{endOfElement(x,i)}$	90%	50%	70%	70%
$\boldsymbol{isCapitalized(x,i)}$ $\boldsymbol{isUppercased(x,i)}$	80%	50%	100%	60%
$\boldsymbol{posNoun(x,i)}$ $\boldsymbol{posVerb(x,i)}$ $\boldsymbol{posAdjective(x,i)}$ $\boldsymbol{posAdverb(x,i)}$ $\boldsymbol{posProperNoun(x,i)}$ $\boldsymbol{posPreposition(x,i)}$	70%	60%	70%	50%
$\boldsymbol{properNoun(x,i)}$	60%	70%	50%	50%
$\boldsymbol{tokenLength(x,i)}$	60%	60%	30%	50%
$\boldsymbol{dataElement(x,i)}$	20%	60%	10%	20%
$\boldsymbol{capitalizedFrequency(x,i)}$	20%	50%	70%	80%

All features contributed to the best performing combination, for all entity types, in at least two of the cross-validation folds. The features which were used the least for the best overall results, $dataElement(x,i)$ and $capitalizedFrequency(x,i)$, were often used when evaluated on the results of the individual entity types. Therefore we believe that all features should be used when applying this approach to other data sets.

We can also observe that the features that detected the names of the entities were always used in the overall results. And, in addition, the features $token(x,i)$, $startOfElement(x,i)$, $endOfElement(x,i)$, $isCapitalized(x,i)$, and $isUppercased(x,i)$ were used very often. This seems to indicate that textual patterns were very relevant for providing evidence for NER.

In the results of the feature $dataElement(x,i)$, it is worth noting that it was used only in 10% or 20% of the folds in the overall results for locations and organizations, but for persons it was used in 60% of the folds. This indicates that the textual patterns where persons are referenced were distinct across data elements, while for the other entity types the patterns were more uniform across data elements. Our analysis pointed that, in the data elements for creators and contributors, the names for persons often appeared in inverse order (that is, *surname, first_names*), while in the other elements they appeared in direct order (that is, *first_names surname*). We therefore conclude that the semantic context given by the data structure is generally not required to allow the recognition of the entities, but in some cases, it can provide importance evidence for the predictive model.

5 Conclusion and Future Work

We presented an approach for the task of named entity recognition in structured data containing free text as the values of its elements. This approach is based on the extraction of features from the text, which allows the predictive model to operate with more independent of lexical evidence than named entity recognition systems developed for grammatically well-formed text.

Our approach was able to achieve a maximum precision of 0.91 at 0.55 recall, and a maximum recall of 0.82 at 0.77 precision. The achieved results were significantly higher than those obtained with the baseline. We believe this level of quality in named entity recognition allows the use of this approach to support a wide range of information extraction applications in digital library metadata.

Although we have specifically studied metadata from the cultural heritage sector, we believe our approach has general applicability to any poorly structured data model.

In future work we will explore the use of ontologies for creating features to improve the recognition of named entities. We will also address the resolution of the recognized named entities in linked data contexts and ontologies.

References

1. Seth, G.: Unstructured Data and the 80 Percent Rule: Investigating the 80%. Technical report, Clarabridge Bridgepoints (2008)
2. Shilakes, C., Tylman, J.: Enterprise Information Portals. Merrill Lynch Report (1998)
3. Sarawagi, S.: Information Extraction. Found. Trends Databases 1, 261–377 (2008)
4. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes 30 (2007)
5. McCallum, A., Freitag, D., Pereira, F.: Maximum entropy Markov models for information extraction and segmentation. In: International Conference on Machine Learning (2000)
6. Martins, B., Borbinha, J., Pedrosa, G., Gil, J., Freire, N.: Geographically-aware information retrieval for collections of digitized historical maps. In: 4th ACM Workshop on Geographical Information Retrieval (2007)
7. Freire, N., Borbinha, J., Calado, P., Martins, B.: A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records. In: ACM/IEEE Joint Conference on Digital Libraries (2011)
8. Sporleder, C.: Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass 4(9), 750–768 (2010)
9. King, P., Poulovassilis, A.: Enhancing database technology to better manage and exploit Partially Structured Data. Technical report, University of London (2000)

10. Williams, D.: Combining Data Integration and Information Extraction. PhD thesis, University of London (2008)
11. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science (2011) ISBN 978-0956599315
12. Michelson, M., Knoblock, C.: Creating Relational Data from Unstructured and Ungrammatical Data Sources. Journal of Articial Intelligence Research 31, 543–590 (2008)
13. Guo, J., Xu, G., Cheng, X., Li, H.: Named Entity Recognition in Query. In: 32nd Annual ACM SIGIR Conference (2009)
14. Du, J., Zhang, Z., Yan, J., Cui, Y., Chen, Z.: Using Search Session Context for Named Entity Recognition in Query. In: 33rd Annual ACM SIGIR Conference (2010)
15. Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. In: Proc. International Conference on Computational Linguistics (1996)
16. Bennett, R., Hengel-Dittrich, C., O'Neill, E., Tillett, B.B.: VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files. In: 72nd IFLA General Conference and Council (2006)
17. Vatant, B., Wick, M.: Geonames Ontology (2006), `http://www.geonames.org/ontology/`
18. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc. (2001)
19. Wallach, H.: Conditional Random Fields: An Introduction. Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania (2004), `http://www.cs.umass.edu/~wallach/technical_reports/wallach04conditional.pdf`
20. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph. 3(4), 235–244 (1990)
21. McCallum, A.: MALLET: A Machine Learning for Language Toolkit (2002), `http://mallet.cs.umass.edu`
22. The Unicode Consortium: Unicode Text Segmentation (2010), `http://www.unicode.org/reports/tr29/`
23. Sekine, S., Isahara, H.: IREX: IR and IE Evaluation project in Japanese. In: Proc. Conference on Language Resources and Evaluation (2000)
24. Michie, D., Spieglhalter, D.J., Taylor, C.C.: Machine learning, neural and statistical classification. Prentice Hall, Englewood Cliffs (1994)
25. Goodman, J.: Sequential Conditional Generalized Iterative Scaling. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 9–16 (2002)
26. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: 43rd Annual Meeting of the Association for Computational Linguistics (2005)
27. Sang, T.K., Erik, F., De, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Conf. on Natural Language Learning (2003)
28. Kohavi, R., John, G.: Wrappers for feature selection. Artificial Intelligence 97(1-2), 273–324 (1997)