NEAT - Named Entities in Archaeological Texts

Gennaro Nolano, Maria Pia di Buono, and Johanna Monti

UniOR NLP Research Group
{gnolano,mpdibuono,jmonti}@unior.it

Abstract. In this document, we propose a pipeline to develop NEAT, a dataset for named entity recognition in archaeological texts based on Europeana's data collections. Our approach relies on a three-step procedure, which includes a semantic projection and the integration of domain terminologies and thesauri. As proof-of-concept, we test the methodology on a small set of data.

Keywords: Europeana Archaeology Collection · Named Entity Recognition · Semantic Projection · Terminology Integration.

1 Introduction

The need for annotated datasets to develop Machine Learning (ML) and Natural Language Processing (NLP) downstream tasks is higly increasing, as proved by several efforts in a lot of domains, mainly legal and medicine domains (REF). Still, some domains, e.g., Cultural Heritage (CH), and some languages suffer from the lack of dedicated datasets which may affect the development of ML and NLP applications.

In order to contribute to full exploit the large amount of available contents, we propose a methodology for the development of Named Entieties in Archaeological Texts (NEAT), an annotated dataset based on Europeana's collection data.

The dataset aims at providing a reliable source of data together with Named Entities (NEs) annotations for the Archaeological domain. As highlithed by [BBR14], automatic NE Recognition (NER) is an important pre-processing step in tasks such as information extraction, question answering, automatic translation, data mining, speech processing and biomedicalscience. Also, it serves as a pre-processing step for deeper linguistic processing such as syntactic or semantic parsing,and co-reference resolution.

Despite the fact that several datasets have been made available for NEs, mainly with reference to shared tasks, e.g., CoNLL 2003 [SD03], and some tools and off-the-shelf libraries[1] provide information on NEs, there is still the need to improve these datasets by integrating domain-specific information, e.g., terminologies,

[1] Among those, the infrastructure CLARIN proposes a set of available tools for several languages. See https://www.clarin.eu/resource-families/tools-named-entity-recognition

thesauri and ontologies. In fact, as far as domain-specific NER is concerned, the aim of dectecting and classifying proper name spans in text may be extended to dectet and classify other entities, identified as representative of the domain. In other words, in many specific domains of knowledge some occurrences, phrases and word sequences, as well as terminology, are representative of conceptualizations which may improve NLP and ML applications in that domain. Thus, as some of these structures may form lists of compounds and terms that have specific meanings only when used with reference to that domain, they could be included and annotated as named entities, contributing to outline formalizations in that specific domain [Buo17].

2 Background

To develop our dataset we refer to previous scholars devoted to both Named Entity Recognition (NER) and semantic projection.
For reasons of space, we mention only some of the researches in the vast literature of these fields.
NER is the task of tagging entities in an unstructured text. One of the possibility of this task is to connect the entities found to a knowledge base (e.g. DBpedia or Wikidata). An example of this interlinking process can be found in [YPF19]. Recently, [MMS19] proposed the use of interlinking to knowledge bases in order to build a dataset for NER tasks.
The use of ontologies for the classification of domain-specific linguistic units has furthermore been proven to be successful, in particular by inducing linguistics and semantic features [Buo17].
Semantic projection can be defined as the process of inducing semantic information from one richer/structured source a poorer/unstructured target. This task has seen a long literature of being used with parallel corpora in multilingual environments, as in [YNW01] and [PL05].

3 Methodology

The methodology we propose relies on the integration of information from several sources, e.g., the Europeana Data Collection, human-annotated texts in the domain of archaeology and a standard thesaurus, i.e., the Italian ICCD Thesaurus[2], one of the best practices adopted by the Italian Ministry of Cultural Heritage (MiBAC).

In order to do so, we define a pipeline composed of the following phases:

1. Off-the-shelf NER (OTS);
2. Semantic projection (SP);
3. Terminology integration (TI).

[2] https://github.com/ICCD-MiBACT/Standard-catalografici/blob/master/strumenti-terminologici/beni%20archeologici/ICCD_Thesaurus_definizione%20del%20bene_reperti%20archeologici.rdf

Data Collection In order to prove the validity of our approach, we select a first set of Europeana's Collection data and proceed with the creation of NEAT mock dataset.
We collect data trough the Europeana API, selecting all items according to two criteria: the provider's country and the Europeana's collection domain. In particular, our query featured the values `europeana_aggregation_edm_country`:Italy and `collection`: `archaeology`, so that we restrict our contents to both a specific domain and a specific country of provenance. Furthermore, we will focus on texts written in Italian language.

Then, starting from the obtained records, we randomly extract 1000 items. For each of the items extracted, from the available metadata we consider only the following information in reference to their content:

- `dcDescription`;
- `dcDescriptionLangAware`;
- `dcLanguageLangAware`;
- `dcTitleLangAware`;
- `edmConcept`.

Among the selected 1000 items, 101 present no value for both `dcDescription` and `dcDescriptionLangAware` fields, and as such have been discarded as they do not have the descriptive text useful for our dataset.
While the remaining items all have values for the description metadata, not all of the texts are present in Italian. In particular, 21 items present only an English description and as such have been casted away.
The final number of items selected for the mock dataset is thus 878 (Table 4), with a total of 67358 tokens.

Table 1. Number of items in the mock dataset.

Formerly Selected Item	1000
Description N/A	101
EN Description	21
Total IT Items	**878**

Off-The-Shelf NER The first step performed is a PoS tagging by means of Tint (The Italian NLP Tool)[3], a standard PoS tagger for Italian based on StanfordCoreNLP[4]. This tool encompasses several NLP modules [CMT18], e.g. a Named Entity Recognizer, a POS tagger and a Lemmatizer. Furthermore, Tint provides Universal Dependencies (UDs) which represent a useful source of syntactic and semantic information for improving annotation projection [].

[3] `https://dh.fbk.eu/research/tint/`
[4] `https://stanfordnlp.github.io/CoreNLP/`

Semantic Projection The semantic information we want to provide in our dataset are more fine-grained than those usually tagged by off-the-shell-tools (i.e. person, organization and location classes). Thus, we refer to the CIDOC Conceptual Reference Model (CRM), a high-level ontology to enable information integration for cultural heritage data and their correlation with library and archive information [Doe03].
We select 11 classes among the ones proposed by the CIDOC CRM, which are considered being informative enough to describe an item (RIF). To project semantic annotation based on those classes, we perform the annotation of a small set of Italian archaeological texts.
Such an annotation task have been carried out within the framework of a larger national project[5], which aims at improving the multilingual access to CH content. We create a small bilingual (IT-EN) and parallel corpus, formed of several types of texts, e.g., brochures, institutional and informative flyers, produced by Italian galleries, libraries, archives, and museums, to promote their archaeological sites and items.

Before performing the annotation task, we proceed with the development of annotationa guidelines (RIF) and a first round of double-annotation on a small set of data, i.e., EN: 21.608 IT: 20.323 tokens, to compute the inter-annotator agreement.

Finally, the annotation has been performed by one trained annotator, In total, we annotate 77.581 tokens with 9192 occurrences (Table 2).

Table 2. Manual Annotated Data

Type	**#Unique**	**#Total**
E1 CRM Entity	85	111
E2 TEMPORAL	5	5
E4 PERIOD	221	830
E5 EVENT	80	170
E18 PHY THING	1525	5248
E28 CON THING	43	81
E39 ACTOR	497	1330
E41 APPELLATION	17	21
E52 TIME SPAN	24	81
E53 PLACE	384	1244
E54 DIMENSION	44	56
E57 MATERIAL	44	185
Total	**2969**	**9192**

Some of the classes are underepresented, e.g., E2 Temporal, E52 TimeSpan, so we might consider either to discard them or to merge them with an upper-level

[5] Semantic Multilingual Access to Cultural Heritage (SMACH) `https://bit.ly/2NOnmIy`

class, e.g., E52 Time Span could be merged with E2 Temporal[6]. Starting from the annotated data, we project the annotation on NEAT dataset, considering that domain-specific occurrences present two levels of representation, which are separated but interlinked. Those levels are: a conceptual-semantic level, pertaining to the knowledge domain and its ontology; and a syntactic-semantic level, pertaining to sentence and word production [Buo17]. Starting from this consideration, we relie on both semantic annotation and syntactic dependencies to project the annotation from our annotated set to the new data in NEAT.

Terminology Integration As already, mentioned, in order to integrate domain-specific terminologies, we refer to the ICCD thesaurus. This resource has been partially formalized into rdf and made available as Linked Open Data [Fel+15]. For each one of the selected 1059 terms, ICCD LOD include an explicit reference to a `skos:closeMatch` linking to to the LOD version of the Getty Art & Architecture Thesaurus (AAT)[7]. The Getty ATT reference allows to extend the conceptual reference for each item, by means of taxonomic domain information, as such a reference is also included into Europeana's collection description metadata (i.e., `edm:Concept`).

4 NEAT Dataset

As already mentioned, NEAT mock dataset contains 878 entries and for each of them we provide several levels of information:

- PoS and UD information;
- Named Entities, including MWE annotation;
- Concept description, at the level of entries.

As mentioned already, the data were collected from Europeana's Collection through the Europeana API. This API takes a query in the form of an url as an input, and as an output it gives a json file containing information about maximum 100 items in the collection at a time. The option `cursor` was used to iterate through all the results found.
We designed a simple Python iterator to navigate from one results page to the next, while collecting for each item the needed information. For this experiment we stopped the process after 10 iterations.
The information were thus downloaded and collected in a json file. We filtered out any item without an `it` entry in their `dcDescriptionLagAware` field. The Italian texts for descriptions were then collected in a single text file, with a `START` token and `END` token appended respectively at the start and at the end of each description.
This text file was processed with the TINT application through its command

[6] Due to the lack of space we do not present the full annotation process with its guidelines and Inter Annotator Agreement scores. This description will be made available in the related project report.

[7] `https://www.getty.edu/research/tools/vocabularies/aat/about.html`

line. The output result is a list of sentences, with each token tagged with its POS and NER, lemmatized and annotated with a representation of dependencies. The results of the NER module were used as a first off-the-shelf annotation.
This corpus thus obtained was semantically enriched in relation to CH domain by projecting a corpus of Italian archaeological texts annotated with a subgroup of the CIDOC CRM, and by integrating domain-specific terminologies through either the ICCD thesaurus for Italian language, or the Getty ATT for English language. A first experiment using a simple exact matching search showed poor results in relation to multi-word expressions (MWEs), and as such we devise an algorithm to extract possible domain-specific phrases in the text. The algorithm is summarized in Figure 1, but for reasons of space some parts were taken out of this explaination.
Once a single-word entity is found in a sentence, we regard it as the head of a potential MWE, and the word is then used to populate a *list of dependencies*: any word governed by the formerly extracted head is written as a new element of the MWE. Then, any dependent word found in this way is similarly used to repopulate the *list of dependencies* by extracting the words depending on it, and the process is repeated once again by using the newly extracted dependent words. The loop stops once no new dependency can be found.

While this first algorithm showed some interesting results, some issues were still present in the MWEs. The main sources of noise were found to be punctuation marks, digits, homonyms (in particular grammatical) and phrases beginning with prepositions or articles. Examples of the most frequent sources of noise can be found in Table 3.

Different solutions are employed to improve the results of this algorithm. While digits and preopositions/articles can simply be removed once the MWE is recreated, punctuation marks and homonyms pose different problems.
Punctuation, for instance, increases the length of the MWE, sometimes across the whole sentence. As such, removing it would not be enough as there would still be words in the MWE that are not actually related to the head. A better solution is to halt the loop once it reaches a punctuation mark, whethever it might be, leaving the MWE as it was before the mark was reached.
Homonyms, on the other hand, cannot be removed as they are usually the head of the MWE. The solution we propose is to check the tagged pos of the token, and if it is tagged as a noun it can be considered the head of the MWE, otherwise it cannot.[8]

Finally, once the MWEs are collected, for each entity found through the ICCD thesaurus, we integrate the domain-specific information in the form of its Getty hierarchy code, which was retrieved by using the `skos:closeMatch` field in the term entry.

[8] While this solution works in our work, we do realize that it cannot be applied in case of domain-specific verb. In this case a dictionary of pos together with the entity would be needed to solve this ambiguity.

Algorithm 1: Algorithm to extract annotated MWEs

input : A list of sentences annotated with Universal Dependencies; a list of annotated entities

output: A list of the entities present in the corpus

```
begin
    max_length ← 4;
    win ← 4;
    list_mwes ← [] for sentence in corpus do
        n←length (sentence);
        for word in sentence do
            if word in annotated_entities then
                e_type←entityType (word);
                mwe← [index (word)];
                anchor←token_index;
                words_to_check←sentence[anchor−win:anchor+win];
                while length(mwe)≤max_length do
                    for w in words_to_check do
                        if governor(w) == anchor then
                            append (index (w)), mwe) ;
                            anchor←index (w);
                            words_to_check←sentence[anchor−win:anchor+win];
                             repeat;
                        else
                            break
                    mwe←sort (mwe);
                    mwe←getWordFromId (mwe);
                    append ([mwe, e_type], list_mwes)
    return list_mwes
```

Source of noise	Example
Punctuation marks	[**')'**, **'-'**, 'sale', 'Vaticane', '-']
Digits	[**'16'**, '-', 'Leonardo']
Homonyms	['uomo', 'vestito', 'da', 'contadino', **'porta'**, 'sulle', 'spalle']
Prepositions/articles at the start of mwe	[**'di'**, **'un'**, 'antico', 'tempio']

Table 3. Sources of noise.

For each step of our pipeline, we increase the number of annotated NEs, as shown in Table 4.

Table 4. NEAT Entities

Type	OTS	TI	SP	Total
E2 TEMPORAL	N/A	-	1	1
E4 PERIOD	N/A	-	57	57
E5 EVENT	N/A	-	71	71
E18 PHY THING	N/A	418	2952	3370
E28 CON THING	N/A	-	17	17
E39 ACTOR	1899(PER+ORG)		1135	3034
E41 APPELLATION	N/A	3	-	3
E52 TIME SPAN	N/A	-	-	-
E53 PLACE	978(LOC)	18	544	1540
E54 DIMENSION	N/A	-	5	5
E57 MATERIAL	N/A	25	384	409
Total	**2877**	**464**	**5166**	**8507**

5 Conclusion

In this proposal, we describe the pipeline to develop a dataset for NER starting from Europeana Archaeology Collection data. To test our approach, we develop NEAT, a mock dataset which encompasses a fine-grained NE classification based on the use of CIDOC CRM classes. NEAT mock dataset has been developed on a small set of items, nevertheless the pipeline, relying on the use of standard thesauri and linked open data from reliable providers, can be extendend to other languages with a small effort.
As recalled by [Law+10], the performance of many machine learning systems is heavily determined by the size and quality of the data used as input to the training algorithms, the final dataset will include about 5000 items in Italian and English. The size of the dataset has been set considering the data required in development of neural network-based approaches for NLP applications.
Both the process of semantic projection and noise reduction have to be improved to guarantee a high precision in the proposed annotations.

References

[YNW01] David Yarowsky, Grace Ngai, and Richard Wicentowski. "Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora". In: *Proceedings of the First International Conference on Human Language Technology Research*. 2001. URL: `https://www.aclweb.org/anthology/H01-1035`.

[Doe03] Martin Doerr. "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata". In: *AI magazine* 24.3 (2003), pp. 75–75.

[SD03] Erik F Sang and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In: *arXiv preprint cs/0306050* (2003).

[PL05] Sebastian Padó and Mirella Lapata. "Cross-linguistic projection of role-semantic information". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 2005, pp. 859–866.

[Law+10] Nolan Lawson et al. "Annotating large email datasets for named entity recognition with mechanical turk". In: *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. 2010, pp. 71–79.

[BBR14] Darina Benikova, Chris Biemann, and Marc Reznicek. "NoSta-D Named Entity Annotation for German: Guidelines and Dataset." In: *LREC*. 2014, pp. 2524–2531.

[Fel+15] Achille Felicetti et al. "Integrating Terminological Tools and Semantic Archaeological Information: the ICCD RA Schema and Thesaurus." In: *EMF-CRM@ TPDL*. 2015, pp. 28–43.

[Buo17] Maria Pia di Buono. "An Ontology-Based Method for Extracting and Classifying Domain-Specific Compositional Nominal Compounds". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2017, pp. 83–88.

[CMT18] Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini. "Tint 2.0: an All-inclusive Suite for NLP in Italian". In: *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it*. Vol. 10. 2018, p. 12.

[MMS19] Daniel Menezes, Ruy Milidiú, and Pedro Savarese. "Building a massive corpus for named entity recognition using free open data sources". In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE. 2019, pp. 6–11.

[YPF19] Beyza Yaman, Michele Pasin, and Markus Freudenberg. "Interlinking SciGraph and DBpedia Datasets Using Link Discovery and Named Entity Recognition Techniques". In: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Ed. by Maria Eskevich et al. Vol. 70. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 15:1–15:8. ISBN: 978-3-95977-105-4. DOI: 10.4230/OASIcs.LDK.2019.15. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10379.