Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities

Francesca Bonin
SCSS and CLCS,
Trinity College Dublin,
Ireland

Fabio Cavulli
University of Trento,
Italy

Aronne Noriller
University of Trento,
Italy

Massimo Poesio
University of Essex, UK,
University of Trento,
Italy

Egon W. Stemle
EURAC,
Italy

Abstract

Developing content extraction methods for Humanities domains raises a number of challenges, from the abundance of non-standard entity types to their complexity to the scarcity of data. Close collaboration with Humanities scholars is essential to address these challenges. We discuss an annotation schema for Archaeological texts developed in collaboration with domain experts. Its development required a number of iterations to make sure all the most important entity types were included, as well as addressing challenges including a domain-specific handling of temporal expressions, and the existence of many systematic types of ambiguity.

1 Introduction

Content extraction techniques – so far, mainly used to analyse news and scientific publications – will play an important role in digital libraries for the humanities as well: for instance, certain types of browsing that content extraction is meant to support, such as entity, spatial and temporal browsing, could sensibly improve the quality of repositories and their browsing. However, applying content extraction to the Humanities requires addressing a number of problems: first of all, the lack of large quantities of data; then, the fact that entities in these domains, additionally to adhering to well established standards, also include very domain-specific ones.

Archaeological texts are a very good example of the challenges inherent in humanities domains, and at the same time, they deepen the understanding of possible improvements content extraction yields for these domains. For instance, archaeological texts could benefit of temporal browsing on the basis of the temporal metadata extracted from the content of the publication (as opposed to temporal browsing based on the date of publication), more than biological publications or general news. In this paper, we discuss the development of a new annotation schema: it has been designed specifically for use in the archaeology domain to support spatial and temporal browsing. To our knowledge this schema is one of only a very few schemata for the annotation of archaeological texts (Byrne et al., 2010), and Humanities domains in general (Martinez-Carrillo et al., 2012) (Agosti and Orio, 2011). The paper is structured as follows. In Section 2 we give a brief description of the corpus and the framework in which the annotation has been developed; in Section 3, we describe a first annotation schema, analysing its performance and its weaknesses; in Section 4 we propose a revised version of the annotation schema, building upon the first experience and, in Section 5, we evaluate the performance of the new schema, describing a pilot annotation test and the results of the inter-annotator agreement evaluation.

2 Framework and Corpus Description

The annotation process at hand takes place in the framework of the development of the Portale della Ricerca Umanistica / Humanities Research Portal (PRU), (Poesio et al., 2011a), a one-stop search facility for repositories of research articles and other types of publications in the Humanities. The portal uses content extraction techniques for extract-

Proceedings of the 6th Linguistic Annotation Workshop, pages 134–138,
Jeju, Republic of Korea, 12-13 July 2012.

ing, from the uploaded publications, citations and metadata, together with temporal, spatial, and entity references (Poesio et al., 2011b). It provides access to the Archaeological articles in the APSAT / ALPINET repository, and therefore, dedicated content extraction resources needed to be created, tuned on the specificities of the domain. The corpus of articles in the repository consists of a complete collection of the journal Preistoria Alpina published by the Museo Tridentino di Scienze Naturali. In order to make those articles accessible through the portal, they are tokenized, PoS tagged and Named Entity (NE) annotated by the TEXTPRO[1] pipeline (Pianta et al., 2008). The first version of the pipeline included the default TEXTPRO NE tagger, EntityPro, trained to recognize the standard ACE entity types. However, the final version of the portal is based on an improved version of the NEtagger capable of recognising all relevant entities in the APSAT/ALPINET collection (Poesio et al., 2011b; Ekbal et al., 2012)

3 Annotation Schema for the Archaeological Domain

A close collaboration with the University of Trento's "B. Bagolini" Laboratory, resulted in the development of an annotation schema, particularly suited for the Archaeological domain, (Table 1). Differently from (Byrne et al., 2010), the work has been particularly focused on the definition of specific archaeological named entities, in order to create very fined grained description of the documents. In fact, we can distinguish two general types of entities: *contextual entities*, those that are part of the content of the article (as PERSONs, SITEs, CULTUREs, ARTEFACTs), and *bibliographical entities*, those that refer to bibliographical information (as PubYEARs, etc.) (Poesio et al., 2011a).

In total, domain experts predefined 13 entities, and also added an *underspecification* tag for dealing with ambiguity. In fact, the archaeological domain is rich of polysemous cases: for instance, the term 'Fiorano' refers to a CULTURE, from the Ancient Neolithic, that takes its name from the SITE, 'Fiorano', which in turn is named from Fiorano Modenese; during the first annotation, those references were decided to be marked as underspecified.

[1] http://textpro.fbk.eu/

NE type	Details
Culture	Artefact assemblage characterizing a group of people in a specific time and place
Site	Place where the remains of human activity are found (settlements, infrastructures)
Artefact	Objects created or modified by men (tools, vessels, ornaments)
Ecofact	Biological and environmental remains different from artefacts but culturally relevant
Feature	Remains of construction or maintenance of an area related with dwelling activities (fire places, post-holes, pits, channels, walls, ...)
Location	Geographical reference
Time	Historical periods
Organization	Association (no publications)
Person	Human being discussed in the text (Otzi the Iceman, Pliny the Elder, Caesar)
Pubauthor	Author in bibliographic references
Publoc	Publication location
Puborg	Publisher
Pubyear	Publication year

Table 1: Annotation schema for Named Entities in the Archaeology Domain

3.1 Annotation with the First Annotation Schema and Error Analysis

A manual annotation, using the described schema, was carried out on a small subset of 11 articles of Preistoria Alpina (in English and Italian) and was used as training set for the NE tagger; the latter was trained with a novel active annotation technique (Vlachos, 2006), (Settles, 2009). Quality of the initial manual annotation was estimated using qualitative analyses for assessing the representativeness of the annotation schema, and quantitative analyses for measuring the inter-annotator agreement. Qualitative analyses revealed lack of specificity of the entity TIME and of the entity PERSON. In fact, the annotation schema only provided a general TIME entity used for marking historical periods (as *Mesolitic, Neolithic*) as well as specific dates (as *1200 A.D.*) and proposed dates(as *from 50-100 B.C.*), although all these instances need to be clearly distinguished in the archaeological domain. Similarly, PERSON had been used for indicating general persons belonging to the document's contents and scientists working on the same topic (but not addressed as bibliographical references). For the inter-annotator agreement on the initial manual annotation, we calculated a kappa value of 0.8, which suggest a very good agreement. Finally, we carried out quantitative analyses of the

NE Type	Details
Culture	Artefact assemblage characterizing a group of people in a specific time and place
Site	Place where the remains of human activity are found (settlements, infrastructures)
Location	Geographical reference
Artefact	Objects created or modified by men (tools, vessels, ornaments, ...)
Material	Found materials (steel)
AnimalEcofact	Animal remains different from artefacts but culturally relevant
BotanicEcofact	Botanical remains as trees and plants
Feature	Remains of construction or maintenance related with dwelling activities (fire places, post-holes)
ProposedTime	Dates that refer to a range of years hypothesized from remains
AbsTime	Exact date, given by a C-14 analysis
HistoricalTime	Macro period of time referring to time ranges in a particular area
Pubyear	Publication year
Person	Human being, discussed in the text (Otzi the Iceman, Pliny the Elder, Caesar)
Pubauthor	Author in bibliographic references
Researcher	Scientist working on similar topics or persons involved in a finding
Publoc	Publication location
Puborg	Publisher
Organization	Association (no publications)

Table 2: New Annotation Schema for Named Entities in the Archaeology Domain

automatic annotation. Considering the specificity of the domain the NE tagger reached high performances, but low accuracy resulted on the domain specific entities, such as SITE, CULTURE, TIME (F-measures ranging from 34% to 70%) In particular SITE, LOCATION, and CULTURE, TIME, turned out to be mostly confused by the system. This result may be explained by the existence of many polysemous cases in the domain, that annotators used to mark as underspecified.

This cross-error analysis revealed two main problems of the adopted annotation schema for Archaeological texts: *1)* the lack of representativeness of the entity TIME and PERSON, used for marking concurrent concepts, *2)* the accuracy problems due to the existence of underspecified entities.

4 A Revised Annotation Schema and Coding Instructions

Taking these analyses into consideration, we developed a new annotation schema (Table 2): the aforementioned problems of the previous section were solved and the first schema's results were outperformed in terms of accuracy and representativeness. The main improvements of the schema are:

1. New TIME and PERSON entities
2. New decision trees, aimed at overcoming underspecification and helping annotators in ambiguous cases.
3. New domain specific NE such as: material
4. Fine grained specification of ECOFACT: AninmalEcofact and BotanicEcofact.

Similarly to (Byrne, 2006), we defined more fine grained entities, in order to better represent the specificity of the domain; however, on the other hand, we also could find correlations with he CIDOC Conceptual Reference Model (Crofts et al., 2011). [2]

4.1 TIME and PERSON Entities

Archaeological domain is characterized by a very interesting representation of time. Domain experts need to distinguish different kinds of TIME annotations.

In some cases, C-14 analysis, on remains and artefacts, allow to detected very exact dating; those cases has been annotated as AbsTIME. On the other hand, there are cases in which different clues, given by the analysis of the settlements (technical skills, used materials, presence of particular species), allow archaeologists to detect a time frame of a possible dating. Those cases have been annotated as *ProposedTime* (eg. *from 50-100 B.C*).

Finally, macro time period, such as *Neolithic, Mesolithic*, are annotated as HistoricalTIME: interestingly, those macro periods do not refer to an exact range of years, but their collocation in time depends on cultural and geographical factors.

4.2 Coding Schema for Underspecified Cases

In order to reduce ambiguity, and helping coders with underspecified cases, we developed the following decision trees:

[2]The repertoire of entity types in the new annotation scheme overlaps in part with those in the CIDOC CRM: for instance, AbsTime and PubYears are subtypes of E50 (Date), HistoricalTime is related to E4 (Period), Artefact to E22 (Man Made Object), etc.

SITE vs LOCATION: coders are suggested to mark as LOCATION only those mentions that are clearly geographical references (eg. *Mar Mediterraneo*, Mediterranean Sea); SITE has to be used in all other cases (similar approach to the GPE markable in ACE); CULTURE vs TIME:
a) coders are first asked to mark as HistoricalTIME those cases in which the mention belongs to a given list of macro period (such as Neolithic, Mesolithic):

- *eg.: nelle societa' Neolitiche (in Neolithic societies).*

b) If the modifier does not belong to that list, coders are asked to try an insertion test: *della cultura + ADJ, (of the ADJ culture)* :

- *lo Spondylus e' un simbolo del Neolitico Danubiano = lo Spondylus e' un simbolo della cultura Neolitica Danubiana (the Spondylus is a symbol of the Danubian Neolithic = the Spondylus is a symbol of the Danubian Neolithic culture).*
- *la guerra fenicia != la guerra della cultura dei fenici (Phoenician war != war of the Phoenician culture).*

Finally, cases in which tests *a)* and *b)* fail, coders are asked to mark and discuss the case individually.

5 Inter-Annotator Agreement and Evaluation

To evaluate the quality of the new annotation schema, we measured the inter-annotator agreement (IAA) achieved during a first pilot annotation of two articles from Preistoria Alpina. The IAA was calculated using the kappa metric applied on the entities detected by both annotators, and the new schema reached an overall agreement of 0.85. In Table 3, we report the results of the IAA for each NE class. Interestingly, we notice a significant increment on problematic classes on SITE and LOCATION, as well as on CULTURE. [3]

Annotators performed consistently demonstrating the reliability of the annotation schema. The new entities regarding coordinates and time seem also to be well defined and representative.

NE Type	Total	Kappa
Site	50	1.0
Location	13	0.76
Animalecofact	3	0.66
Botanicecofact	6	-0.01
Culture	4	1.0
Artefact	18	0.88
Material	11	0.35
Historicaltime	6	1.0
Proposedtime	0	NaN
Absolutetime	0	NaN
Pubauthor	48	0.95
Pubyear	32	1.0
Person	2	-0.003
Organization	7	0.85
Puborg	0	NaN
Feature	36	1.0
Publoc	2	-0.0038
Coordalt	0	NaN
Geosistem	0	NaN
Datum	2	1.0

Table 3: IAA per NE type: we report the total number of NE and the kappa agreement.

6 Conclusions

In this study, we discuss the annotation of a very specific and interesting domain namely, Archaeology: it deals with problems and challenges common to many other domains in the Humanities. We have described the development of a fine grained annotation schema, realized in close cooperation with domain experts in order to account for the domain's peculiarities, and to address its very specific needs. We propose the final annotation schema for annotation of texts in the archaeological domain. Further work will focus on the annotation of a larger amount of articles, and on the development of domain specific tools.

Acknowledgments

This study is a follow up of the research supported by the LiveMemories project, funded by the Autonomous Province of Trento under the Major Projects 2006 research program, and it has been partially supported by the 'Innovation Bursary' program in Trinity College Dublin.

[3] Five classes are not represented by this pilot annotation test; however future studies will be carried out on a significantly larger amount of data.

References

M. Agosti and N. Orio. 2011. The cultura project: Cultivating understanding and research through adaptivity. In Maristella Agosti, Floriana Esposito, Carlo Meghini, and Nicola Orio, editors, *Digital Libraries and Archives*, volume 249 of *Communications in Computer and Information Science*, pages 111–114. Springer Berlin Heidelberg.

E. J. Briscoe. 2011. Intelligent information access from scientific papers. In J. Tait et al, editor, *Current Challenges in Patent Information Retrieval*. Springer.

P. Buitelaar. 1998. CoreLex: Systematic Polysemy and Underspecification. Ph.D. thesis, Brandeis University.

K. Byrne and E. Klein, 2010. Automatic extraction of archaeological events from text. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology*, Williamsburg, VA

K. Byrne, 2006. Proposed Annotation for Entities and Relations in RCAHMS Data.

N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. 2011. Definition of the CIDOC Conceptual Reference Model. ICOM/CIDOC CRM Special Interest Group, 2009.

A. Ekbal, F. Bonin, S. Saha, E. Stemle, E. Barbu, F. Cavulli, C. Girardi, M. Poesio, 2012. Rapid Adaptation of NE Resolvers for Humanities Domains using Active Annotation. In *Journal for Language Technology and Computational Linguistics (JLCL) 26 (2)*:39–51.

A. Herbelot and A. Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 165–174, Stroudsburg, PA, USA.

A. Herbelot and A. Copestake 2010. Underquantification: an application to mass terms. In *Proceedings of Empirical, Theoretical and Computational Approaches to Countability in Natural Language*, Bochum, Germany, 2010.

B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, and R. Sprugnoli. I-CAB: the Italian Content Annotation Bank: pages 963–968.

A.L. Martinez Carrillo, A. Ruiz, M.J. Lucena, and J.M. Fuertes. 2012. Computer tools for archaeological reference collections: The case of the ceramics of the iberian period from andalusia (Spain). In *Multimedia for Cultural Heritage*, volume 247 of *Communications in Computer and Information Science*, Costantino Grana and Rita Cucchiara, editors, pages 51–62. Springer Berlin Heidelberg.

M. Palmer, H. T. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

E. Pianta, C. Girardi, and R. Zanoli. 2008. The textpro tool suite. In *Proceedings of 6th LREC*, Marrakech.

M. Poesio, P. Sturt, R. Artstein, and R. Filik. 2006. Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence. In *Discourse Processes* 42(2): 157-175, 2006.

M. Poesio, E. Barbu, F. Bonin, F. Cavulli, A. Ekbal, C. Girardi, F. Nardelli, S. Saha, and E. Stemle. 2011a. The humanities research portal: Human language technology meets humanities publication repositories. In *Proceedings of Supporting Digital Humanitites (SDH)*, Copenhagen.

M. Poesio, E. Barbu, E. Stemle, and C. Girardi. 2011b. Structure-preserving pipelines for digital libraries. In *Proceedings of LaTeCH*, Portland, OR.

J. Pustejovsky. 1998. The semantics of lexical underspecification. *Folia Linguistica*, 32(3-4):323?348.

J. Pustejovsky and M. Verhagen. 2010. Semeval-2010 task 13 : Evaluating events, time expressions, and temporal relations. *Computational Linguistics*, (June 2009):112–116.

B. Settles. 2009. Active learning literature survey. *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison.

A. Vlachos. 2006. Active annotation. In *Proceedings EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento.