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

Francesca Bonin[1], Fabio Cavulli[2], Aronne Noriller[2], Massimo Poesio[2 3], Egon W. Stemle[4]
[1]Trinity College Dublin, Ireland [2]University of Trento, Italy [3]University of Essex, UK, [4]EURAC, Italy

Summary

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; in close collaboration with Humanities scholars, we discuss an annotation schema for Archaeological texts. Its development required a number of iterations to make sure all the most important entity types were included, as well as to address domain specific challenges (handling of temporal expressions, and the existence of many systematic types of ambiguity).

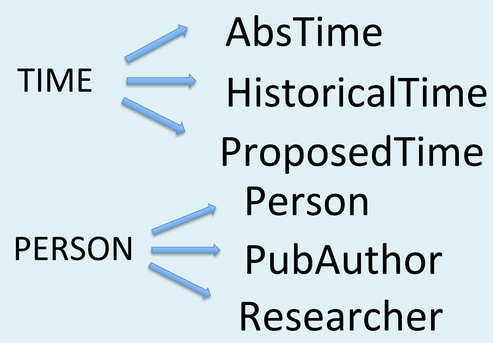
Framework

The annotation process at hand takes place in the framework of the development of the Portale della Ricerca Umanistica/Humanities Research Portal (PRU)[1], a one-stop search facility for repositories of research articles and other types of publications in the Humanities. The portal uses content extraction techniques to automatically extract citations and metadata including temporal, spatial, and entity references from the publications being uploaded. It provides access to the Archeological articles in the APSAT/ALPINET repository, hence domain dedicated resources needed to be created.

A Revised Annotation Schema and Coding Instructions: 2nd step

Main improvement of the schema:

1. New TIME and PERSON entities:



2. New decision trees, aimed at overcoming underspecification and helping annotators in ambiguous cases as:
   - SITE vs LOCATION
   - CULTURE vs TIME
3. New domain specific NE such as: MATERIAL, CoordAlt (geo coordinates).
4. Fine grained specification of ECOFACT: AninmalEcofact and BotanicEcofact.

Annotation Schema for the Archaeological Domain: 1st step

First annotation schema for archaeological texts:

- **Contextual entities**: PERSON, SITE, CULTURE, ARTEFACTs ECOFACT, LOCATION, FEATURE, TIME, ORGANIZATION.
- **Bibliographical entities**: PubYEARs, PubORG, PubAUTHOR, PubLOC.
- Underspecification tag: ambiguity cases

A manual annotation and an automatic annotation were carried out with this annotation scheme.

An error analysis on the two annotation revealed:

1. **Lack of representativeness of the entity TIME and PERSON, used for marking concurrent concepts,**
2. **Accuracy problems due to the existence of underspecified entities.**
3. **Inter Annotator Agreement (IAA) 80%**

Final NE Types for the Archaeological domain

Culture	Artefact
Site	Material
Location	Feature
BotanicEcofact	ProposedTime
AnimalEcofact	AbsTime
PubYear	HistoricalPeriod
PubLoc	Person
PubAuthor	Researcher
PubOrg	Organization

Evaluation new schema and IAA

- First pilot annotation to evaluate the quality of the new annotation schema.
- Inter-annotator agreement (IAA): calculated using the kappa metric.
- **IIA: 85%.**
- A significant increment on problematic classes can be noticed: **SITE (Kappa 100%), LOCATION (Kappa 76%), CULTURE (Kappa 100%).**

Conclusions

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.

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