Rapid Adaptation of NE Resolvers for Humanities Domains via Active Annotation

Asif Ekbal[1] Francesca Bonin[2] Sriparna Saha[1]
Egon Stemle[2] Eduard Barbu[2] Fabio Cavulli[2]
Christian Girardi[3] Massimo Poesio[24]

[1]IIT Patna

[2]University of Trento

[3]Fondazione Bruno Kessler

[4]University of Essex

ACRH

Named Entities in Humanities Domains

Many entities mentioned in scholarly articles in subjects such as Archeology, History, or History of Art are not among the types most studied in Computational Linguistics.
E.g., in the Archaeology texts we studied in this work, the most frequent entities after TIME and LOCATION are

- ECOFACTs (remains of animals or plants found on a site)
- SITEs
- ARTEFACTs

In order to recognize such entities in text, NE Recognizers have to be retrained on newly annotated material.
BUT collections of humanities material tend to consist of many different domains of small size (and funding for annotation very limited)

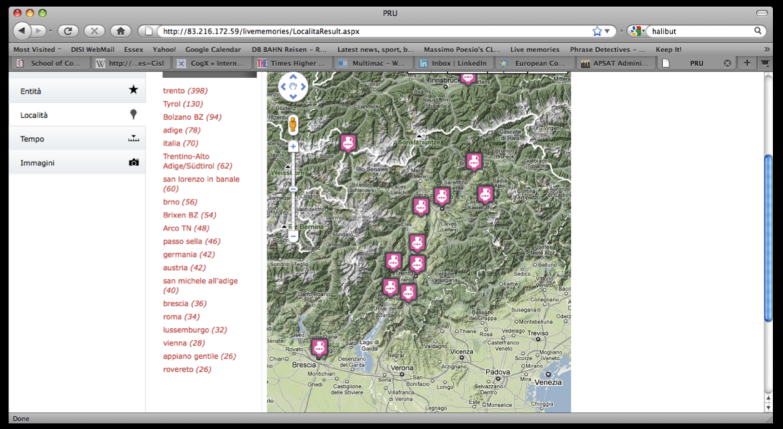
Minimizing Work through Active Annotation

- ACTIVE LEARNING techniques (Settles 2009), already used for NE tagging in the biomedical domain by Vlachos (2006), appear ideally suited for the task of creating data to retrain NE taggers with minimal effort.
- In this work we used active learning to annotate NEs in a corpus of scholarly articles in Archeology in support of the creation of the Portale Ricerca Umanistica del Trentino.

Il Portale Ricerca Umanistica Trentina

www.portalericercaumanistica.org

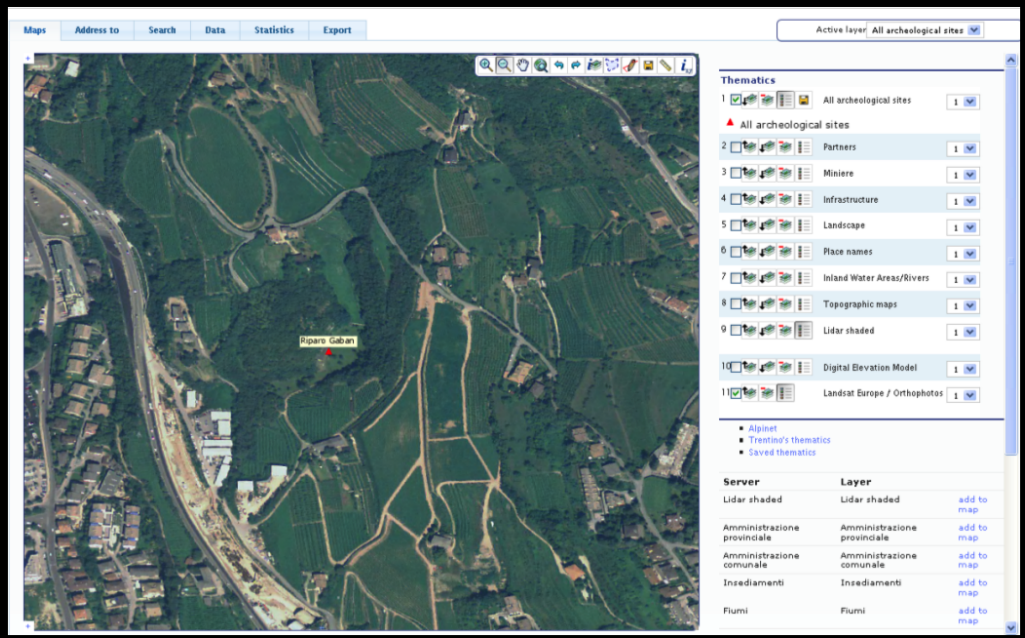
- A one-stop search facility for repositories of (multilingual) scholarly articles in the Humanities held by digital libraries, museums and archives Trentino
- Information extraction techniques used to extract information about entities, spatial locations, and temporal locations
- used to allow ENTITY-BASED, SPATIAL-BASED, and TEMPORALLY-BASED BROWSING
- First repository to be made accessible: ALPINET / APSAT



The Alpinet / APSAT Repository

http://alpinet.mpasol.it/webgis/

- A pilot SPATIAL HUMANITIES project developed by the University of Trento's B. BAGOLINI ARCHAEOLOGY LAB, allowing scholars to visualize archaeological sites in the Alps through a WEB GIS interface
- Through the portal, scholars can also access archaeological ARTICLES about a site through the WEB GIS interface
- Among the holdings: complete collection of PREISTORIA ALPINA



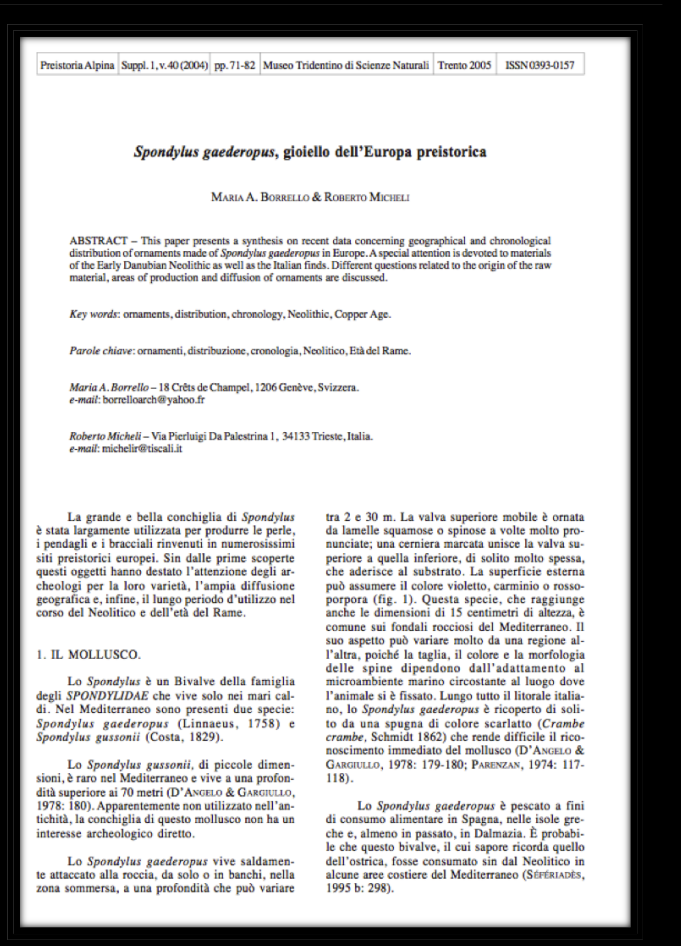
Entities in Preistoria Alpina

Named Entity 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, cimiteries, production site, ...)
ARTEFACT	Objects created or modified by men (tools, vessels, ornaments, ...)
ECOFACT	Biological and environmental remains different from artefacts but culturally relevant (e.g., *Spondylus*)
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 (e.g., Ötzi the Iceman, Pliny the Elder, Caesar)
PUBAUTHOR	author in bibliographic references
PUBLOC	publication location
PUBORG	publisher
PUBYEAR	publication year

(approximately 24% of entities belong to the new types)

A Structure-Sensitive, Multilingual Pipeline

- The articles to be browsed through the PRU are processed by a pipeline that processes the text to extract semantic indices (Poesio et al, LaTeCH 2011)
- The pipeline is based on the TEXTPRO pipeline (Pianta et al LREC 2008) but has two distinguishing features:
  1. It is STRUCTURE SENSITIVE
  2. It is CONSTITUENT-LEVEL MULTILINGUAL



Preistoria Alpina | Suppl. 1, v. 40 (2004) | pp. 71-82 | Museo Tridentino di Scienze Naturali | Trento 2005 | ISSN 0393-0157

***Spondylus gaederopus*, gioiello dell'Europa preistorica**

MARIA A. BORRELLO & ROBERTO MICHELI

ABSTRACT – This paper presents a synthesis on recent data concerning geographical and chronological distribution of ornaments made of *Spondylus gaederopus* in Europe. A special attention is devoted to materials of the Early Danubian Neolithic as well as the Italian finds. Different questions related to the origin of the raw material, areas of production and diffusion of ornaments are discussed.

Key words: ornaments, distribution, chronology, Neolithic, Copper Age.

Parole chiave: ornamenti, distribuzione, cronologia, Neolitico, Età del Rame.

Maria A. Borrello – 18 Crêts de Champel, 1206 Genève, Svizzera.
e-mail: borrelloarch@yahoo.fr

Roberto Micheli – Via Pierluigi Da Palestrina 1, 34133 Trieste, Italia.
e-mail: michelir@tiscali.it

La grande e bella conchiglia di *Spondylus* è stata largamente utilizzata per produrre le perle, i pendagli e i bracciali rinvenuti in numerosissimi siti preistorici europei. Sin dalle prime scoperte questi oggetti hanno destato l'attenzione degli archeologi per la loro varietà, l'ampia diffusione geografica e, infine, il lungo periodo d'utilizzo nel corso del Neolitico e dell'età del Rame.

1. IL MOLLUSCO.

Lo *Spondylus* è un Bivalve della famiglia degli *SPONDYLIDAE* che vive solo nei mari caldi. Nel Mediterraneo sono presenti due specie: *Spondylus gaederopus* (Linnaeus, 1758) e *Spondylus gussonii* (Costa, 1829).

Lo *Spondylus gussonii*, di piccole dimensioni, è raro nel Mediterraneo e vive a una profondità superiore ai 70 metri (D'ANGELO & GARGIULLO, 1978: 180). Apparentemente non utilizzato nell'antichità, la conchiglia di questo mollusco non ha un interesse archeologico diretto.

Lo *Spondylus gaederopus* vive saldamente attaccato alla roccia, da solo o in banchi, nella zona sommersa, a una profondità che può variare tra 2 e 30 m. La valva superiore mobile è ornata da lamelle squamose o spinose a volte molto pronunciate; una cerniera marcata unisce la valva superiore a quella inferiore, di solito molto spessa, che aderisce al substrato. La superficie esterna può assumere il colore violetto, carminio o rosso-porpora (fig. 1). Questa specie, che raggiunge anche le dimensioni di 15 centimetri di altezza, è comune sui fondali rocciosi del Mediterraneo. Il suo aspetto può variare molto da una regione all'altra, poiché la taglia, il colore e la morfologia delle spine dipendono dall'adattamento al microambiente marino circostante al luogo dove l'animale si è fissato. Lungo tutto il litorale italiano, lo *Spondylus gaederopus* è ricoperto di solito da una spugna di colore scarlatto (*Crambe crambe*, Schmidt 1862) che rende difficile il riconoscimento immediato del mollusco (D'ANGELO & GARGIULLO, 1978: 179-180; PARENZAN, 1974: 117-118).

Lo *Spondylus gaederopus* è pescato a fini di consumo alimentare in Spagna, nelle isole greche e, almeno in passato, in Dalmazia. È probabile che questo bivalve, il cui sapore ricorda quello dell'ostrica, fosse consumato sin dal Neolitico in alcune aree costiere del Mediterraneo (SÉFÉRIADÈS, 1995 b: 298).

Adapting the NE Tagger to a New Domain

- The objective of this work was to develop methods for rapid adaptation of a NE tagger to a new domain

Active Learning

- In traditional random sampling, unlabelled data are chosen for annotation at random.
- In active learning, the most useful data are carefully selected for annotation on the basis of their INFORMATIVENESS:
  - A classifier is initially trained on a small set of SEED items.
  - This first classifier is then used to label previously unlabelled items.
  - Some of these items are identified as MOST INFORMATIVE and given to human coders to label
  - The most informative items are added to the training data and the process is iterated.

Our Informativity Criterion

- Many informativity criteria have been proposed
- The simplest: choose items on which PROBABILITY IS LOWEST
  - Often DOESN'T WORK VERY WELL
- Alternative: choose items on which DIFFERENCE AMONG PROBABILITY OF TOP TWO LABELS (a measure of uncertainty) is lowest
- Both methods require a classification method that can assign a probability to items

Conditional Random Fields

- An undirected graphical model (Lafferty et al, 2001)
- Used to compute the conditional probability of values on OUTPUT NODES given values on INPUT NODES
- A special case are conditionally-trained probabilistic finite state automata
- Can incorporate large numbers of non-independent features
- Are becoming the preferred method for NE tagging

Active Annotation with CRF

0: TRAIN a first classifier on a (small) set of training data.

1: Evaluate the system on the gold standard test data.
2: Test on the development data and calculate the conditional probabilities of all the output classes.
3: Compute the CONFIDENCE INTERVAL (CI) between the two most probable classes for each token.
4: If CI is below the THRESHOLD VALUE (set to 0.1 and 0.2) then
4.1: Add the NE token along with its sentence identifier and CI in a list of EFFECTIVE SENTENCES, selected for active annotation (named as ES).
5: Sort ES in ascending order of CI.
6: Select the top most 10 sentences.
7: Remove the 10 sentences along with the preceding one and following one sentences from the development set.
8: Add the sentences to the training set.
9: RETRAIN the CRF classifier.
10: Repeat steps 1-9 until the performance in two consecutive iterations is the same.

Dataset

To test the method, 11 articles from *Preistoria Alpina* for a total of around 50K words were annotated by the authors according to the scheme discussed previously. These 11 articles were broken up as follows:

Set	# docs	# tokens	# NEs
Training	5	20,739	2,611
Additional training	3	5,292	622
Test	3	11,534	1,582

Features

LANGUAGE-INDEPENDENT features (cfr. Ekbal and Saha, 2010, 2011)

- SURFACE WORD FEATURES: suffix / prefix, word length, capitalization, presence of digits
- CONTEXTUAL FEATURES: previous and following words, presence of content words, first in sentence
- LEXICAL FEATURES: POS, lemma, infrequency, word normalization cluster

LANGUAGE-DEPENDENT features

- GAZETTEER (the list of entities in the ALPINET / APSAT database)

Experimental Design

Planned comparisons:

1. ACTIVE ANNOTATION vs. RANDOM SAMPLING
2. Different THREESHOLDS (0.1 vs. 0.2)
3. Different NUMBER OF EXTENSION SENTENCES (10 vs. 30)
4. With / Without a GAZETTEER

Results: Random vs. Active Selection (two thresholds)

Iteration number	Threshold=0.1			Threshold=0.2			Baseline (random)		
	r	p	F	r	p	F	r	p	F
1	63.02	65.48	64.23	64.32	67.83	66.03	64.64	66.35	65.47
2	64.73	67.11	65.90	65.84	68.81	67.29	64.21	65.99	65.09
3	65.08	67.92	66.47	66.10	69.6	67.81	65.40	66.90	66.14
4	65.66	68.41	67.01	66.80	70.09	68.41	65.86	67.73	66.78
5	66.82	69.62	68.19	67.68	70.92	69.27	65.54	67.25	66.39
6	67.31	70.06	68.66	68.26	70.26	69.24	65.66	67.25	66.44
7	67.63	70.31	68.94	68.26	70.54	69.38	65.77	67.41	66.58
8	67.63	70.31	68.94	68.26	70.54	69.38	66.90	68.56	67.72
9	67.86	70.57	69.19	68.83	70.99	69.89	67.19	68.90	68.04
10	67.86	70.57	69.19	68.83	70.99	69.89	67.19	67.90	68.04

Results: With Gazetteer

Iteration number	Threshold=0.1					Threshold=0.2				
	r	p	F	#s add	#NE add	r	p	F	#s add	#NE add
1	67.51	66.93	67.18	27	113	65.52	68.93	67.18	27	113
2	66.08	67.29	65.65	23	115	66.08	69.29	67.65	23	115
3	66.46	69.36	67.88	24	118	66.46	69.36	67.88	24	118
4	67.29	70.08	68.66	25	123	67.29	70.08	68.66	25	123
5	68.87	71.24	70.04	19	68	68.87	71.24	70.04	19	68
6	69.19	71.19	70.18	8	16	68.86	71.57	70.19	17	35
7	69.19	71.19	70.18	1	3	69.51	71.47	70.48	3	5
8	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0
9	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0
10	69.19	71.19	70.18	0	0	69.51	71.47	70.48	0	0

Results: Summary

- Active selection (F=69.89%) almost two points better than random selection (F=68.04%)
- Threshold = 0.2 (F = 69.89%) better than Threshold = 0.1 (F=69.19%)
- Using gazetteers (F=70.48%) leads to another small improvement

Error Analysis

- Analyzed
  - BOUNDARY ACCURACY (see paper)
  - CLASSIFICATION ACCURACY

Classification accuracy

Class	TP	FP	FN	Tot Retr	Total	P	R	F-M
B-Artefact	26	70	21	96	47	0.27	0.55	0.36
B-Culture	12	34	17	46	29	0.26	0.41	0.32
B-Ecofact	164	37	107	201	271	0.82	0.61	0.69
B-Feature	0	9	0	9	-	0	-	
B-Location	117	78	52	195	169	0.6	0.69	0.64
B-Person	0	20	0	20	-	0	-	
B-Pubauthor	380	23	55	403	435	0.94	0.87	0.91
B-Publoc	2	1	3	3	5	0.67	0.4	0.5
B-Puborg	1	0	7	1	8	1	0.13	0.22
B-Pubyear	265	20	10	285	275	0.93	0.96	0.95
B-Site	57	64	66	121	123	0.47	0.46	0.47
B-Time	97	14	44	111	141	0.87	0.69	0.77
I-Artefact	70	76	27	146	97	0.48	0.72	0.58
I-Culture	20	48	26	68	46	0.29	0.43	0.35
I-Ecofact	232	40	121	272	353	0.85	0.66	0.74
I-Feature	0	0	14	0	14	-	0	-
I-Location	262	164	66	426	328	0.62	0.8	0.69
I-Person	0	0	24	0	24	-	0	-
I-Pubauthor	64	9	40	73	104	0.88	0.62	0.72
I-Publoc	6	0	30	6	36	1	0.17	0.29
I-Puborg	13	1	24	14	37	0.93	0.35	0.51
I-Pubyear	0	0	2	0	2	-	0	-
I-Site	168	98	95	266	263	0.63	0.64	0.64
I-Time	400	40	66	440	466	0.91	0.86	0.88
Total	2356	817	946	3173	3302	-	-	-
O	11703	126	38	11829	11741	0.99	1	0.99

Classification Accuracy and Underspecification

- The NE tagger obtains very accurate results for categories like `Pub-year`, `Pub-author`, and `Time`, but not so good with `Artefact`, `Culture` and `Site`
- These categories are difficult for coders as well:
  - The classes `Culture` and `Site` are systematically correlated as the culture is named from a so-called TYPE SITE (e.g., *Starcevo*):
  - As a result, 55% of `Culture` NEs are correctly identified, but 20% are marked as `Site`

Conclusions

- Annotation does lead to better results than random sampling;
- We can achieve reasonable results with relatively small amounts of trained data
- Ongoing work is focused on improving the coding scheme for the first domain and retraining the NE tagger
- Future work includes testing the generality of our results by incorporating a new domain

An Extended Set of Entities

Named Entity Type	Details	
Spatial Entities	PUBLOC	publication location
	GEO COORDINATES	53 degrees Latitude, 62 longitude
	SITE	Place where the remains of human activity are found (settlements, infrastructures, cimiteries, production site, ...)
	LOCATION	other geographical reference
Temporal Entities	PUBYEAR	publication year
	MACRO-PERIOD	*Neolitico*
	TEMPORAL-INTERVAL	*dal 50 al 100 a.C.*
	EXACT-TIME	*nel 45 a. C.*
Persons	PUBAUTHOR	author in bibliographic references
	PERSON	human being discussed in the text (e.g., Ötzi the Iceman, Pliny the Elder, Caesar)
Organizations	PUBORG	publisher
	ORGANIZATION	association (not publisher)
Other entities	MATERIALS	*rame*
	ARTEFACT	Objects created or modified by men (tools, vessels, ornaments, ...)
	ECOFACT	Biological and environmental remains different from artefacts but culturally relevant (e.g., *Spondylus*)
	FEATURE	Remains of construction or maintenance of an area related with dwelling activities (fire places, post-holes, pits, channels, walls, ...)
	CULTURE	Artefact assemblage characterizing a group of people in a specific time and place

Thanks for your attention!