Active Annotation for Handling Named Entities in Humanities Domain

Asif Ekbal
Department of Computer Science and Engineering
IIT Patna, India
Email: asif@iitp.ac.in

Joint works with
Ans D Alghamdi, Francesca Bonin, Sriparna Saha
Fabio Cavulli, Massimo Poesio

Outline

- Background
- Why Active Learning?
- Proposed approach
  - Ensemble Method
- Features for Named Entity Recognition
- Dataset & Experiments
- Conclusions

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
- Archaeology texts
  - frequent entities after TIME and LOCATION are

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

    SITEs

    ARTEFACTs
- To recognize such entities in text requires
  - New annotated datasets
  - BUT collections of humanities material tend to consist of many different domains of small size (and funding for annotation very limited)

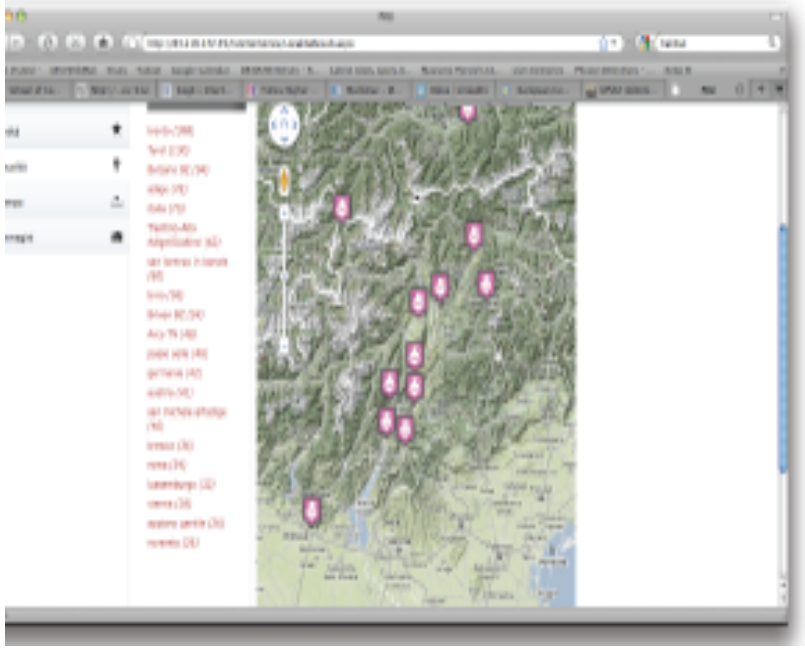
Minimizing Work through Active Annotation

- Active learning techniques (Settles, 2009) is ideally suited for this task
  - Already used for NE tagging in the biomedical domain by Vlachos (2006)

- What has been done here?
  - 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*

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

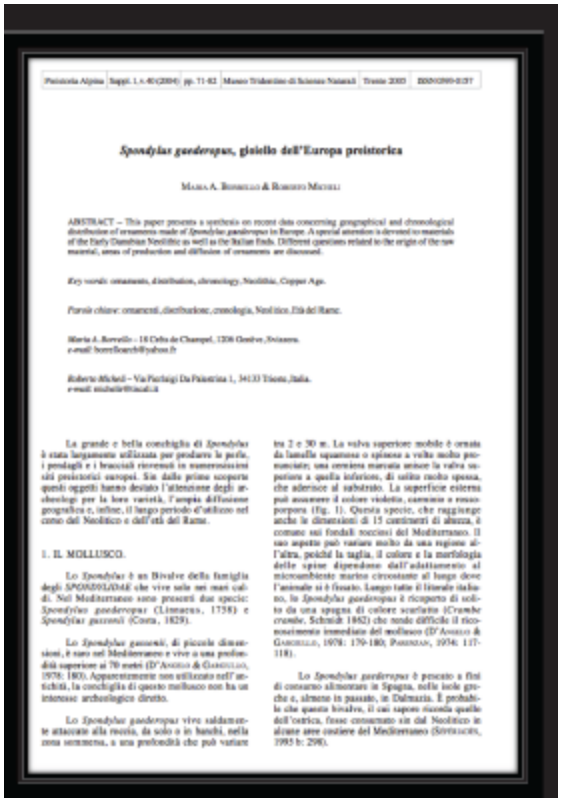
Named Entity Types

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
Material	Found materials
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
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	Scientists working on similar topics or persons involved in a finding
Publoc	Publication location
Puborg	Publisher
Organization	Association (no publications)

A Structure-Sensitive, Multilingual Pipeline

- Articles to be browsed through the PRU are processed by a pipeline that tokenizes, POS-tags, and NE-tags 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:
  - Structure sensitive
  - Constituent-level-multilingual



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 Cebs du Champel, 1206 Genève, Svizzera.
e-mail: borrelloarch@yahoo.fr

Roberto Micheli – Via Pacifici 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 ricercati 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 sulle rive rocciose calde. 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 cresta marcata unisce la valva superiore a quella inferiore, di colito molto sposa, che aderisce al substrato. La superficie esterna può assumere il colore violetto, cremisino e 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 (Stevanović, 1993 b: 298).

Active Learning

- Selects informative samples from large amount of unlabeled examples

- Advantages
  - Active Learning optimizes the control of model growth
  - Leads to drastic reductions in the amount of annotation
  - Greatly reduces the time and costs involved in preparing the data as well as the model
  - Makes more efficient use of the learner's time by asking them to label only instances that are most useful for the trainer

Active Learning

- Most informative examples
  - Instances where classifier is most uncertain
- Restricts the amount of learning by the learning algorithm
- Kind of semi-supervised learning framework
  - But selects most uncertain samples

Active Learning

- Traditional random sampling
  - Unlabelled data are chosen for annotation at random
- Active learning
  - Train a classifier on a small set of SEED items
  - Use this classifier to label unlabelled items
  - Select data carefully on the basis of their *informativeness*
  - Most informative instances are sent to the coders for feedbacks
  - Add the informative items to the training data
  - Repeat the process

Our Selection Criterion (Informativeness measure)

- Exists many selection criterion criteria
  - *The simplest*: Choose items on which prediction probability is lowest
  - Often doesn't work very well
- *Alternative (margin sampling)*: Choose items on which differences among prediction probabilities of top two labels (a measure of uncertainty) is lowest
  - *Hypothesis*: items for which this difference is smaller are those of which the classifier is less certain
- Both methods require a classification method that can assign a probability/confidence score to items

Proposed Approach

- Model-Ensemble Framework

- Ensemble
  - Combination of more than one classifier
  - Improves generalization
  - Effective when diverse classifiers are combined

- Ensemble in the present work
  - Support Vector Machine (SVM)
  - Conditional Random Field (CRF)

Proposed Approach

- Unlabeled data from development set chosen in such a way that performance on test data improves
- SVM
  - For each token of the development set, a SVM classifier produces confidence scores
  - *Confidence*-distance from different separating hyper planes
- CRF
  - For each token of the development set, a CRF classifier produces confidence scores
  - *Confidence*-Marginal probabilities

Proposed Approach

- Normalize the distance values in the range [0,1]- confidence value for a particular class

- Selection criterion
  - Differences between the confidence values of the most probable two classes for a token

Proposed Approach

- Define a threshold on the confidence interval
- Select the uncertain samples for each of SVM and CRF
- For each uncertain sample
  - Select that particular sentence from the development set
- Combine these two different sampled sentences
  - Each sentence has its own confidence scores
  - For the common sentences the confidence scores are set equal to the minimum of two values
- Arrange the sentences according to the descending order

Proposed Approach

- Select the top most 10 sample sentences
- Add these sentences both to the *training as well as development*
  - *Results in increased size of training and development sets*
- Execute the algorithm for the maximum 20 iterations
- Finally select the model that showed highest performance

Basic Steps of the Algorithm

Step 1: Train the base classifiers with the initial training data and evaluate with the gold standard test data.
Step 2: Train the base classifiers with the initial training data and evaluate with the development data.
Step 3: Calculate the confidence value of each token for each output class.
Step 4: Normalize the confidence scores within the range of [0,1].
Step 5: Compute the confidence interval (CI) between the two most probable classes for each token of the development data. This is computed on the outputs of both SVM and CRF.
Step 6: From each of dev_output, perform the following operations:
    Step 6.1: if CI is below the threshold value (set to 0.2) then add the NE token along with its sentence identifier and CI in a set of effective sentences, selected for active annotation.
    Step 6.2: Create two different sets, (*Set SVM* and *Set CRF*) for two classifiers.

Step 7: Combine two sets into one, named as EA in such a way that if the sentence identifiers are same, then for that sentence $CI_{new} = min(CI_{SVM}, CI_{CRF})$. All the dissimilar sentences are added as they are.
Step 8: Sort EA in ascending order of CI_{new}.
Step 9: Select the top most 10 sentences, and remove these from the development data.
Step 10:Add the sentences to the training set. This generates new training set. Retrain the SVM and CRF classifiers and evaluate with the test set.
Step 11:Repeat steps 3-10 for some iteration (10 in our case).

NE Features: domain-independent

- Local contexts : Preceding and succeeding few words
- Word Suffix
  - Not necessarily linguistic suffixes
  - Fixed length character strings stripped from the endings of words
- Word Prefix
  - Fixed length character strings stripped from the beginning of the words

NE Features: domain-independent

- Named Entity Information: Dynamic NE tag (s) of the previous token (s)
- FirstWord (binary valued): First word of the sentence is most probably NEs
- Part of Speech (PoS) information- PoS of the current and/or surrounding token(s)
  - Extracted from TextPro

NE Features: domain-independent

- Chunk information-Chunk of the current and/or surrounding token(s)
  - Extracted from TextPro
- Lemma- Root word of the token
  - Extracted from TextPro
- Unknown token feature-checks whether current token appears in training

NE Features: domain-independent

- Word class feature-Certain kinds of NEs, which belong to the same class, are similar to each other
  - capital letters→ A, small letters→a, number→O and non-English characters→-
  - consecutive same characters are squeezed into one character
  - groups similar names into the same NE class

- Capitalization- checks whether the word starts with a capitalized letter

NE Features: domain-specific

- Gazetteers based feature
  - Feature value is set to 1 or 0 depending upon the presence or absence of a word in the gazetteer
  - Two gazetteers- SITES (2078) and CULTURES (98)
- MultiWordNet based feature
  - Significant low recall observed for *animalecofact*, *botanicecofact* and *artefact*
  - Missed candidates looked into the WordNet and assigned proper class labels
  - Performance further drops when *animalecofact*, *botanicecofact* and *artefact* considered altogether
  - Performance improves when only *animalecofact and botanicecofact* considered

Datasets

- To test the method, 25 articles from Preistoria Alpina were annotated by the authors according to the scheme discussed previously

Set	# documents	#NEs
Training	19	8900
Development	3	694
Test	3	1665

Experimental Design

- Planned comparisons
  - ACTIVE ANNOTATION vs. RANDOM SAMPLING
  - Different THREESHOLDS (0.1 vs. 0.2)
  - With / Without a GAZETTEER (the list of entities in the ALPINET / APSAT database
  - Training with ALL sentences vs Training with only sentences containing NEs

- Collapsed classes (*increases performance approximately by 2%*)
  - Pub-author, Person, Researcher-→ Person
  - Pub-year, Absolute-time → Absolute-time

Results: Random vs. Active Selection (CRF) [Gazetteers: Sites and Cultures]

Iteration	Threshold=0.1			Threshold=0.2			Random		
	R	P	F1	R	P	F1	R	P	F1
1	46.12	76.47	57.54	47.87	78.61	59.51	46.01	76.17	57.37
2	47.23	76.61	58.43	48.51	78.72	60.03	46.68	76.23	57.90
3	47.57	76.66	58.71	48.79	78.69	60.23	46.77	76.29	57.99
4	48.07	77.03	59.19	49.13	79.24	60.65	46.86	76.54	57.20
5	48.18	77.04	59.28	49.16	79.09	60.63	47.02	76.62	58.28
6	48.37	77.23	59.48	49.22	79.31	60.74	47.31	76.74	58.53
7	48.56	77.19	59.62	49.43	79.26	60.89	47.22	76.62	58.43
8	48.71	77.28	59.76	49.77	79.34	61.16	47.14	76.64	58.38
9	48.63	77.23	59.68	49.22	79.15	60.69	47.27	76.76	58.51
10	48.70	77.12	59.70	49.59	79.08	60.95	47.32	76.63	58.51
11	48.67	77.24	59.71	49.55	79.14	60.94	47.27	76.71	58.49
12	48.61	77.11	59.63	49.49	78.78	60.79	47.21	76.54	58.40

Results: Active Selection (CRF) [Gazetteers: Sites and Cultures]

Iteration no	# Sentences selected	#Sentences added	#mentions added
1	95	10	45
2	85	10	54
3	75	10	41
4	62	10	20
5	53	10	22
6	41	10	31
7	39	10	37
8	24	10	26
9	16	10	19
10	6	10	13
11	1	10	1

Results: Active Selection

- Results using MultiWordNet: *animalecofact and botanicecofact*
  - R=52.80 P=77.99 and F=62.97
  - Improvement of 3.03 points in recall, 1.34 points in precision and 1.81 points in F-measure

- Results using MultiWordNet: *animalecofact*, *botanicecofact* and *artefact classes*
  - R=53.23 P=47.91 F=50.42
  - Significant drop in the overall performance (*more than 12% F-measure*)

Results: Active Selection (Class-wise)

Class	Recall	Precision	F-measure
Absolutetime	96.17	89.32	92.46
Artefact	25.38	66.00	36.67
AnimalEcofact	44.56	80.16	57.28
BotanicEcofact	37.50	29.03	32.73
Culture	36.14	71.42	48.00
Site	27.51	58.10	37.34
Feature	40.74	29.72	34.38
HistoricalTime	45.20	85.58	59.16
Location	50.45	79.28	61.66
Material	5.00	5.00	5.00
Organization	0	0	0
Person	79.22	88.83	83.75
ProposedTime	40.71	59.38	48.31

Results: Ensemble of CRF and SVM

Iteration	Threshold=0.2		
	R	P	F1
1	50.95	77.27	61.41
2	52.14	77.42	62.31
3	53.17	78.13	63.28
4	53.24	78.22	63.36
5	54.37	79.34	64.52
6	54.61	79.52	64.75
7	54.65	79.61	64.81
8	54.67	79.64	64.83
9	54.69	79.67	64.86
10	54.66	78.62	64.48
11	54.63	78.63	64.47
12	54.65	78.61	64.47
13	54.65	78.61	64.47

Experimental Designs: Current Setups

- Active expert learning
  - Feedbacks from experts: *entities are selected by experts*
  - Feedbacks from non-experts: *entities are selected by non-experts*
  - Sample selection criterion: *random vs. active learning vs. experts*

- Sample selection criterion
  - Random sample selection
  - Sample selection by active learning (*CRF based*)
  - Sample selection by expert annotators

In all selection mechanisms, 50 entities are selected for feedbacks

Experimental Designs: Current Setups

- Different combinations for feedbacks
  - Expert Annotator-Uncertain Entities (active learning)
  - Expert Annotator-Entities Selected by Experts
  - Non-expert Annotator –Uncertain Entities (active learning)
  - Non-expert Annotator – Entities Selected at Random

For each unlabeled document, maximum 200 entities are selected

Conclusions

- Annotation does lead to better results than random sampling
- We can achieve reasonable results with relatively small amounts of trained data
- Implementation of more features
  - Bag-of-words feature, informative words in contexts, domain-specific features, orthographic features
- Informative sample selection criterion
  - Different margins for different classes
- Retraining with more data

References

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

- Poesio M, Barbu E, Bonin F, Cavulli F, Ekbal A, Girardi C, Nardelli F, Saha S, , Stemle E (2011). The humanities research portal: Human language technology meets humanities publication repositories. In: Proc. of SDH, Copenhagen

- Poesio, Barbu, Stemle, and Girardi] Poesio M, Barbu E, Stemle E, Girardi C (2011) Structure-preserving pipelines for digital libraries. In: Proc. Of LaTeCH, Portland,

Thank you for your attention!