Active Expert Learning for the Digital Humanities

Ans Alghamdi[1,6]*, Francesca Bonin*[2]*, Asif Ekbal*[3]*, Sriparna Saha*[3]*, Fabio Cavulli*[4]*, Sara Tonelli*[5]*, Massimo Poesio*[1,4]*, Udo Kruschwitz*[1]

(1) University of Essex
(2) Trinity College Dublin
(3) Indian Institute of Technology Patna
(4) University of Trento
(5) Fondazione Bruno Kessler
(6) Albaha University

adalgh@essex.ac.uk, boninf@scss.tcd.ie, asif@iitp.ac.in, sriparna@iitp.ac.in, fabio.cavulli@unitn.it, satonelli@fbk.eu, poesio@essex.ac.uk, udo@essex.ac.uk

ABSTRACT

Current platforms for paper sharing among scholars, such as Research Gate, could support *Active Expert Learning*, whereby the paper being uploaded is processed using human language technology techniques, and feedback is asked of the scholar doing the upload using active learning techniques to minimize the amount of feedback requested. We show that this approach could outperform traditional active learning as well as randomly asking for feedback.

KEYWORDS: named entity recognition, active learning, e-science.

1 Background and Motivations

Platforms for sharing paper among scholars, such as Research Gate or the ACL Web, have had a very positive impact on research, making papers available much more easily, in particular to scholars working in disciplines with less resources, such as the Humanities. For maximum profit however new methods are needed both to make it easier for scholars to enter their papers and have them properly indexed, and to facilitate retrieval, in particular to enthusiastic amateurs who may not be aware of the appropriate terminology. One such platform is the APSAT / ALPINET under development as a collaboration between the University of Trento's Archeology Lab and a number of other research centres in the ALPINET network, carrying out research on Alpine archeology. The portal, structured around WebGIS technology and a large database, enables scholars both to share papers and to visualize the sites of excavation discussed in these papers. Uploading these papers however requires scholars to enter a great deal of information (e.g., citation information, sites and cultures mentioned, etc). The ultimate goal of our work (Poesio et al., 2011a) is to develop methods to automatize insofar as possible this process of extracting information from the papers, and also to develop novel visualization techniques allowing non-experts easier access to this information (e.g., spatial / entity / temporal based browsing) in addition to standard keyword search. To this end, we have been developing a pipeline able to extract much of this information automatically (Poesio et al., 2011b).

The key problem in developing such technology is the lack of training data. Each scholarly field requires specialized data (in fact, for e.g., citation extraction, almost every publisher requires slightly different data), but creating large amounts of annotated data for each domain is both very cost-sensitive and time consuming (Laws et al., 2012). In the case of the Humanities, there is the additional problem that the total amount of data is fairly small compared to, e.g., biology or the medical sciences, and the percentage of this in digital form is even smaller. It is therefore imperative to adopt techniques to maximize the training benefit from small amounts of data, such as **active learning** (Settles, 2009). In this paper, we argue that the context of platform for sharing research papers enables a form of active learning that we call **active expert learning**–the provision of feedback to an active learner from the part of the scholar uploading the paper–and that this approach makes better use of the training data than random checking or active learning where the feedback is provided by non-experts. We demonstrated the use of this methodology with named entity recognition, but it is potentially applicable to other aspects of information extraction from scholarly documents as well.

2 Active Expert Learning: The Idea

In active learning, a classifier is trained on a small sample of the data, known as the **seed** examples. The classifier is thereafter applied to a huge pool of unlabeled data to select the informative samples. The informative samples are annotated by the expert(s) and the cycle is repeated. This allows the classifier to refine the decision boundaries between the classes. Using active learning we can reduce the amount of manual annotation which is necessary for creating a training corpus. The strength of active learning is that only a subset of tokens which are useful for a given classifier are selected for annotation, in contrast with the traditional **random sampling** approach, where unlabeled data is selected for annotation at random. (Vlachos, 2006) introduced the term **active annotation** to refer to the methodology of using active learning to drive annotation

In this paper we propose a variant of active annotation that we call **active expert learning**: combining active learning with crowdsourcing the provision of feedback required by active

learning to the very experts who produced the papers being processed.

In this framework, when a scholar uploads a paper in the APSAT / ALPINET, a HLT pipeline (Poesio et al., 2011b) including a named entity recognizer initially trained on a small amount of manually annotated data is used to extract information required by the portal for each paper. The user is then asked to provide feedback on the aspects of the extraction that the named entity recognizer is less certain about. This feedback is then used to retrain the named entity recognizer.

3 Active Learning with CRF-Based Named Entity Recognition

In this Section we briefly introduce the supervised machine learning approach we used to develop our NER system, based on Conditional Random Fields (CRF), and the approach to selecting the most informative samples we adopted in our work.

3.1 Conditional Random Fields

Conditional Random Fields (CRFs) (Lafferty et al., 2001) are one of the dominant paradigms to train models for NER. CRFs calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $s =< s_1, s_2, \ldots, s_T >$ given an observation sequence $o =< o_1, o_2, \ldots, o_T >$ is calculated as:

$$P_\wedge(s|o) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k \times f_k(s_{t-1}, s_t, o, t)),$$

where $f_k(s_{t-1}, s_t, o, t)$ is a feature function whose weight λ_k, is to be learned via training. The values of the feature functions may range between $-\infty, \ldots + \infty$, but typically they are binary. When applying CRFs to the NER problem, an observation sequence is a token of a sentence or document of text and the state sequence is its corresponding label sequence.

3.2 Active Annotation with CRF

The most significant issue in active learning is to determine the samples that will be most effective to the classifier. In the literature (Vlachos, 2006; Settles, 2009) a number of techniques have been proposed for selecting the informative samples from the unlabeled data. Our sample selection criterion is based on margin sampling that is defined based on the conditional probabilities of CRF model. For each token of the unlabeled data, CRF classifier produces the confidence values of each class. In our selection criterion we consider an instance to be uncertain if the difference between the confidence values of the most probable two classes is below some threshold value. We hypothesize this with the assumption that items for which this difference is smaller are those of which the classifier is less certain. We experiment with different threshold values and finally fix to 0.2 based on the results obtained on a held out data. In each iteration of the algorithm we select the effective sentences from the unlabeled data and add to training. As mentioned earlier the selection of instances could vary depending upon the model we choose. The samples selected to be uncertain by the classifier are passed to the users or experts for their feedbacks. After getting the feedbacks the instances are added to the training set. In random selection and expert selection strategy we don't select uncertain samples based on the margin sampling criterion of CRF. However in all the cases the classifier is subjected to re-training after adding new instances. We add only the sentence that contains the most informative example. We iterate the algorithm for the maximum of 20 steps, and always select the CRF model that produces

the highest performance. More details about the method will be provided in the final paper, if accepted.

3.3 Features

We identify and implement the features for the NER task based on different possible combinations of available words, local contextual information, shallow syntactic information and orthographic constructs. We implement the features without using deep domain-specific knowledge and/or resources so that these can be easily adapted to some other domains and applications. For domain specific resources we use of gazetteers of sites and cultures and the MultiWordNet. A full list of the features used will be provided in the final version of the paper, if accepted.

4 Experiments and Results

In this Section we discuss the experimental setting we adopted and the results.

4.1 Experimental design

Our designs were designed to compare two factors: 1) entities' selection method (**+A** for Active Learning, **-A** for random selection) and 2) the nature of the feedback (**+E** for Expert feedback, i.e. archeologist, **-E** for not-expert feedback, i.e. non-archeologists). This results in the following four conditions:

1. *-A -E*, or **Random Selection with Non-Experts**: In this condition, corresponding to annotation as normally carried out, the new unlabeled instances to add to the training data are selected randomly, and feedback on these instances is provided by non-expert users.

2. *+A -E*, or **Active Learning with Non-Expert Feedback**: In this condition, the active learning methods discussed in the previous Section are used to identify uncertain instances, and non-expert users provide feedback. These new instances are thereafter added to the training data with the feedback.

3. *+A +E*, or **Active Expert Learning**: This is the condition that we hypothesize will lead to the best results. In this condition, the active learning methods discussed in the previous Section are used to identify uncertain instances and expert users (i.e. archeologist) provide feedback.

4. *-A +E*, or **Active Learning with Expert Selection**: This condition is meant as an upper bound. In this setup the most problematic cases are identified by the expert(s) themselves.

4.2 Implementation

In order to compare these conditions, a website was set up where users after logging in would be presented with documents to give feedback on. In close collaboration with experts (e.g. archaeologists); each user was classified in one of two categories: expert or non-expert. The documents to give feedback on were processed using four distinct versions of our pipeline. Initially each pipeline includes a NER system based on the CRF model described in the previous Section and trained with the available training data. Then users start providing feedback on new documents; currently each user is asked to provide feedback on 50 entities. As feedback is collected, whenever a given number of documents (4 in this implementation) is completed by subjects accessing the website in a given condition, the platform uses the feedback to retrain the model to be used to tag NEs to be seen in that condition, and the documents are re-tagged.

Condition	Precision	Recall	F-score
Baseline	0.409	0.789	0.538
-A -E	0.491	0.922	0.641
+A -E	0.539	0.879	0.668
+A +E	0.548	0.887	**0.677**
-A +E	0.540	0.883	0.670

Table 1: Active Expert Learning compared to other settings

4.3 Results

In Table 1 we report the results of our experiment, comparing the four scenarios (described in the previous Section) with a baseline, a model trained on the initial data, without any additional information. We evaluate against a set consist of about 7000 annotated tokens. All systems trained with additional data outperform the baseline by a considerable margin in terms of Precision, Recall and F-score —though not significant ($p > 0.5$). Active learning without experts (+A -E) leads to an F-score of $F = 0.668$, and, as expected, the proposed approach, Active Expert Learning (+A +E) outperforms all the other models, including random selection by experts (-A +E), with an F-score of $F = 0.677$.

5 Conclusions

We proposed that a version of active learning in which experts provide feedback is particularly suited for emerging platforms for sharing research papers among scholars; and demonstrated that this approach outperforms other ways of improving the performance of pipelines for information extraction from scholarly documents by adding more data.

References

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Laws, F., Heimer, F., and Schütze, H. (June 3-8, 2012). Active learning for coreference resolution. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–512,, Montreal, Canada. Association for Computational Linguistics.

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

Poesio, M., Barbu, E., Stemle, E., and Girardi, C. (2011b). Structure-preserving pipelines for digital libraries. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 54–62. Association for Computational Linguistics.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

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