Learning More with Less: Reducing Annotation Effort with Active and Interactive Learning*

Shilpa Arora and Manas Pathak
{shilpaa,manasp}@cs.cmu.edu
Advisor: Eric Nyberg

Annotation learning is an important task for many kinds of text analysis. Statistical machine learning techniques have been developed to learn annotation models over labeled data that are then used to annotate unlabeled data. One of the major bottlenecks of a conventional machine learning approach is getting sufficient pre-labeled training data: annotating text is a time consuming, tedious and error prone process. Moreover, as all the training examples are not equally informative or equally easy to annotate, it is beneficial to identify examples that would help the model to converge with minimal user annotation effort. Active learning aims at reducing the amount of labeled training data required for learning the target concept. Interactive learning keeps the user in the loop and seeks to improve the performance of the learner with minimal input from the user.

Active learning consists of an initial set of labeled examples and a large set of unlabeled examples. The classifier is trained on the labeled examples and is used to annotate the unlabeled examples. From the pool of annotated examples, selective sampling is used to create a small subset of examples for the user to label. The labeled data is added to the training pool and this iterative process of training, selective sampling and annotation is repeated until all the data is processed.

In this project, we present a generalized active learning framework for learning text annotations. This work is part of the Interactive Annotation Learning (IAL) project [Nyberg et al., 2007] at LTI. The goal of this project is to develop a generalized framework for learning any type of annotation (simple or structured). This project extends an initial software engineering framework[1] developed for the IAL project, in order to demonstrate active and interactive learning techniques for named-entity recognition. We used the Stanford Named Entity Recognizer [Finkel et al., 2005] along with the Reuters corpus with labeled named entities from the CoNLL 2003 shared task [Sang et al., 2003].

Selective sampling strategies rank and select examples from the unlabeled documents in the pool. We establish measures that can be used to evaluate different selection strategies for active and interactive learning. We analyze and compare the proposed selection criteria in terms of these measures. Pool-based sampling strategies are known to perform better than stream-based strategies which consider each document individually irrespective of the alternatives [McCallum et al., 1998]. We show that requesting the user to label the annotations which the model is least certain about can help to learn the target concept faster, while achieving a performance comparable to a traditional approach [Thompson et al., 1999]. In our work, we investigated the following pool-based selective sampling strategies.

*The work described here is part of a larger research initiative as described in [Nyberg et al., 2007]

[1]Developed by Ben Lambert and José Alavedra as part of SE II project

1. Average annotation confidence

$$AC = \frac{\sum_{i=1}^{n} conf(l_i)}{N}$$

where: $conf(l_i)$ = confidence assigned to annotation l_i

N = number of annotations

2. Relative number of annotations below threshold (average annotation confidence over the training pool)

$$RBT = \frac{t - min_t}{max_t - min_t}$$

and, threshold $th = \frac{\sum_{i=1}^{D} AC_i \, n_i}{\sum_{i=1}^{D} n_i}$

where: D = number of Documents

n_i = number of annotations in document i

t = number of annotations with confidence below threshold th

3. Relative document length

$$RDL = \frac{d - min_d}{max_d - min_d}$$

where: d = number of words

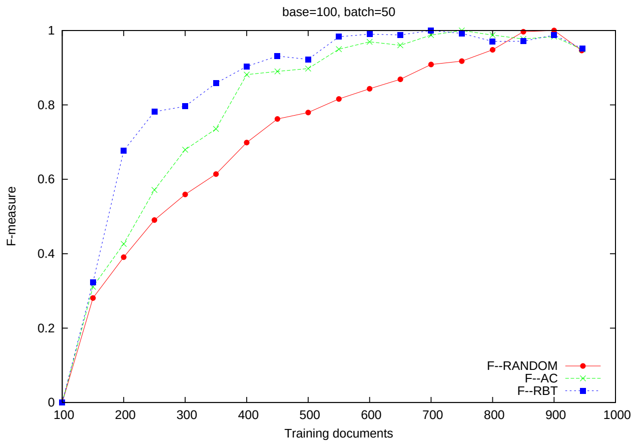
4. Annotation density

$$AD = \frac{\text{\#words in annotations}}{\text{\#words in document}}$$

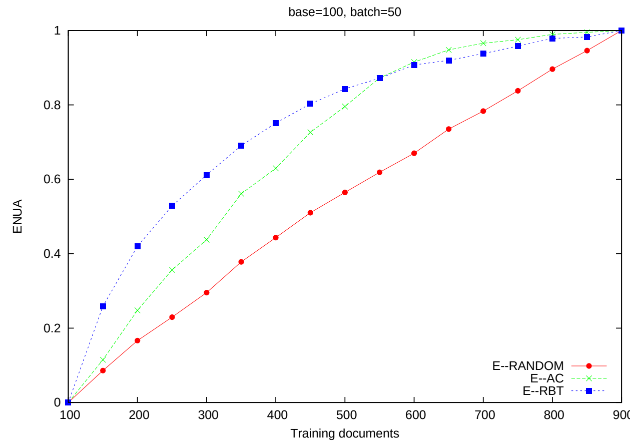
Average annotation confidence and number of annotations below threshold are used to select the documents with least certainty, while document length and annotation density are used to select documents with lower annotation effort. To combine these two strategies we use a weighted-sum approach. The weights are estimated based on heuristics and domain knowledge, similarly to other work on multi-criterion active learning [Shen et al., 2004] [Kim et al., 2006]. In the future, we plan to use machine learning techniques to estimate automatically the optimum values for these weights.

For evaluation, we use two metrics: Precision and recall (F-measure) of annotations for the performance of the named entity recognition task, and Expected Number of User Actions (ENUA) [Kristjannson et al., 2004] for annotation effort. The selection criteria AC and RBT intend to achieve maximal F-measure with fewer examples, while the criteria RDL and AD intend to reduce the average ENUA to achieve comparable convergence. As a baseline, we use a random recommendation strategy that arbitrarily selects the examples to add to the training pool.

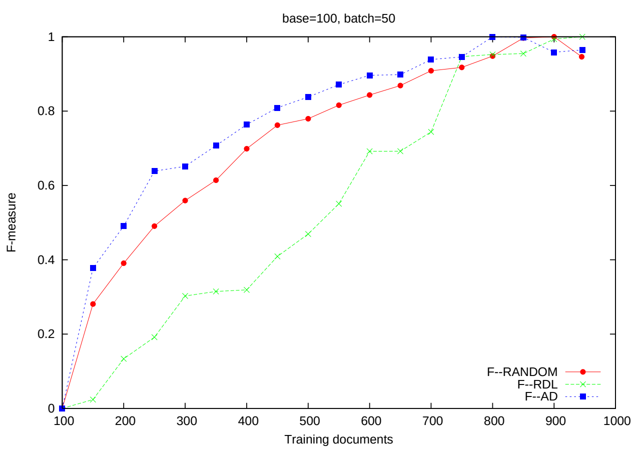
Figs. 1 (a)-(d) show that the confidence-based recommenders (AC & RBT) outperform the random recommender in F-measure but do worse on ENUA. The recommenders based on user interactions (RDL & AD) outperform the random recommender in ENUA but perform worse in F-measure. Fig. 2 shows that a



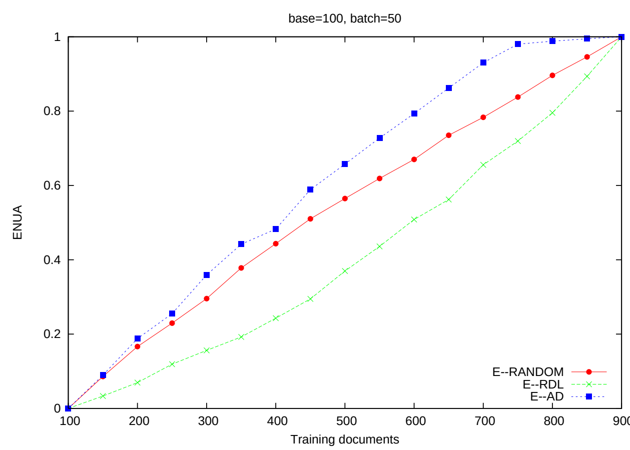
(a) F-measure for AC & RBT (Best: RBT)



(b) ENUA (normalized) for AC & RBT

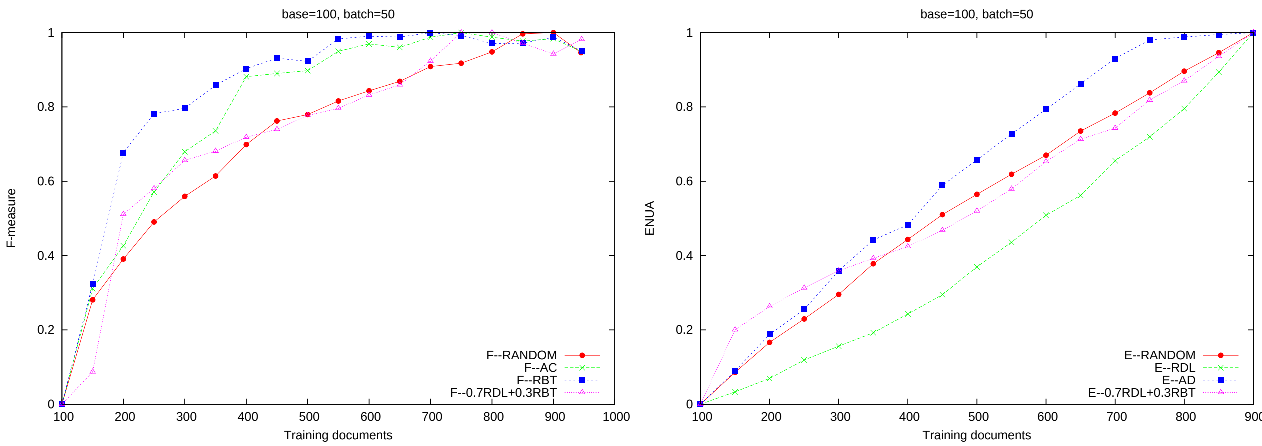


(c) F-measure for RDL & AD



(d) ENUA (normalized) for RDL & AD (Best: RDL)

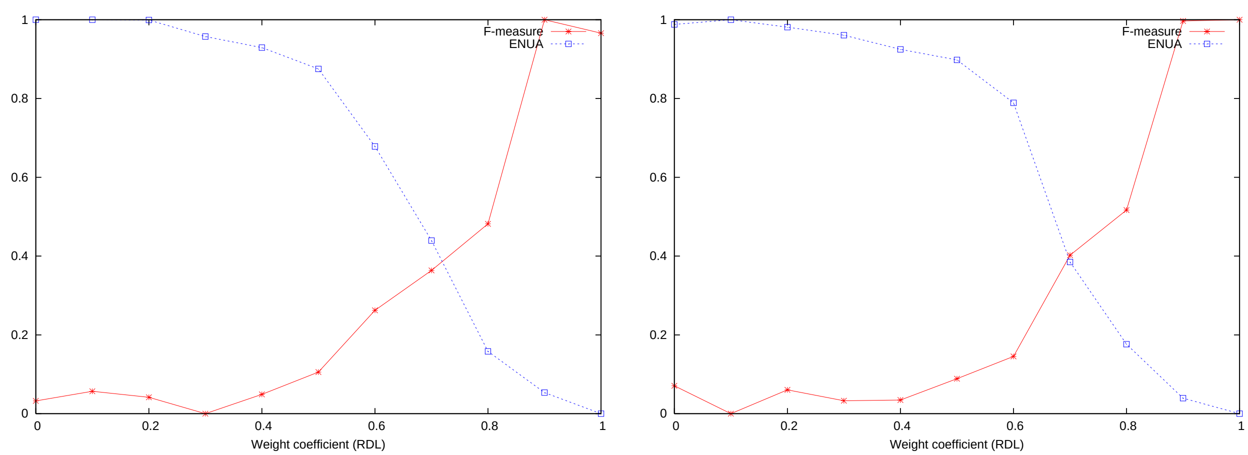
Figure 1: Comparison of performance in terms of F-measure & ENUA for different recommendation strategies.



(a) F-measure

(b) ENUA

Figure 2: Normalized F-measure and ENUA for a combination of RDL and RBT and its comparison with random and best of Active and Interactive measures.



(a) ΔF-measure and ΔENUA for 600 documents

(b) ΔF-measure and ΔENUA at 900 documents

Figure 3: Normalized differences between the optimal and observed values for F-measure and ENUA, using a weighted sum of RBT & RDL as the selection criterion ($D_F = |F_{RBT} - F_C|$ and $D_{ENUA} = |ENUA_{RDL} - ENUA_C|$). The point of intersection indicates the weight value where F-measure is maximum for the minimum value of ENUA.

#Training Documents	F_{RBT}	F_R	F_C	$\Delta_1 F$	$\Delta_2 F$	$ENUA_{RDL}$	$ENUA_R$	$ENUA_C$	$\Delta_1 ENUA$	$\Delta_2 ENUA$
200	0.77	0.71	0.73	0.04	-0.02	4.48	11.41	17.10	-12.62	-5.69
300	0.79	0.74	0.76	0.03	-0.02	8.64	17.16	21.88	-13.24	-4.72
400	0.82	0.77	0.77	0.04	0.00	12.80	23.76	25.10	-12.30	-1.34
500	0.82	0.79	0.78	0.04	0.01	18.90	29.17	29.86	-10.96	-0.69
600	0.83	0.80	0.80	0.04	0.01	25.56	33.87	36.44	-10.88	-2.57
700	0.84	0.82	0.81	0.02	0.00	32.62	38.92	40.90	- 8.28	-1.98
800	0.83	0.83	0.83	0.00	0.00	39.34	43.96	47.20	- 7.86	-3.24
900	0.83	0.84	0.82	0.02	0.02	49.17	48.59	53.63	- 4.46	-5.04

Table 1: A comparison of F-measure & ENUA for selection strategies. Δ_1F & Δ_1ENUA indicate the difference between the combined measure and optimum, while Δ_2F & Δ_2ENUA indicate the difference between the combined measure and random.

combination of a confidence based strategy (best of AC & RBT) and a user interaction based strategy (best of RDL & AD) achieves performance comparable to the best recommender in both F-measure and ENUA. In Table 1, we show a comparison of normalized F-measure and normalized ENUA (% of total operations) for the combined strategy with the other selection strategies. We are currently working to establish the upper bound on ENUA for the optimal confidence based selection strategy, so that we can measure the savings in user effort for different training sets.[1] We are also working to establish optimum weights for the combined selection strategy.

Figs. 3 (a) and (b) plot the difference between optimal and observed F-measure (ΔF) and ENUA ($\Delta ENUA$) when examples are selected according to a weighted combination of RBT & RDL for certain training dataset sizes. The point of intersection between these two curves indicates the weight value where F-measure is maximum for the minimum value of ENUA.

In this work we demonstrate that it is fruitful to investigate this combination of confidence-based and user-effort based strategies for annotation learning. Most of the other work in the literature has focused mainly on minimizing the number of examples required to learn the model and not the amount the user effort. [Kristjannson et al., 2004] use annotation effort as an evaluation measure but to the best of our knowledge, it has not yet been used as a criterion for selective sampling.

References

[Finkel et al., 2005] Finkel J., Grenager T., and Manning C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

[Kim et al., 2006] Kim S., Song Y., Kim K., Cha J., Lee G., MMR-based Active Machine Learning for Bio Named Entity Recognition. In proceedings of HLT-NAACL, 2006.

[Kristjannson et al., 2004] Kristjannson T., Culotta A., Viola P., McCallum A., Interactive information extraction with constrained conditional random fields. In proceedings of AAAI, 2004

[1] These additional results will be available at the time of the SRS presentation.

[McCallum et al., 1998] McCallum, A. K., Nigam, K., 1998. Employing EM in pool-based active learning for text classification. Proceedings of ICML-98, 15th International Conference on Machine Learning (pp. 350358). Madison, US: Morgan Kaufmann Publishers, San Francisco, US.

[Nyberg et al., 2007] Nyberg E., Arora S., Pathak M., Lambert B., Alavedra J., Interactive Annotation Learning : Active Learning for Real-World Text Analysis.(Unpublished Manuscript)

[Sang et al., 2003] Sang E, De Meulder F., 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In proceedings of CoNLL, 2003. (http://www.cnts.ua.ac.be/conll2003/ner)

[Shen et al., 2004] Shen D., Zhang J., Su J., Zhou G., Tan C., Multi-Criteria-based Active Learning for Named Entity Recognition. In proceedings of ACL, 2004.

[Thompson et al., 1999] Thompson, C. A., Califf, M. E., Mooney, R. J., 1999. Active learning for natural language parsing and information extraction. In proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 406-414.