Assessing the Costs of Sampling Methods in Active Learning for Annotation

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, Peter McClanahan
Department of Computer Science
Brigham Young University
Provo, UT 84602, USA
robbie_haertel@byu.edu, ringger@cs.byu.edu, kseppi@cs.byu.edu,
jlcarroll@gmail.com, petermcclanahan@gmail.com

Abstract

Traditional Active Learning (AL) techniques assume that the annotation of each datum costs the same. This is not the case when annotating sequences; some sequences will take longer than others. We show that the AL technique which performs best depends on how cost is measured. Applying an hourly cost model based on the results of an annotation user study, we approximate the amount of time necessary to annotate a given sentence. This model allows us to evaluate the effectiveness of AL sampling methods in terms of time spent in annotation. We acheive a 77% reduction in hours from a random baseline to achieve 96.5% tag accuracy on the Penn Treebank. More significantly, we make the case for measuring cost in assessing AL methods.

1 Introduction

Obtaining human annotations for linguistic data is labor intensive and typically the costliest part of the acquisition of an annotated corpus. Hence, there is strong motivation to reduce annotation costs, but not at the expense of quality. Active learning (AL) can be employed to reduce the costs of corpus annotation (Engelson and Dagan, 1996; Ringger et al., 2007; Tomanek et al., 2007). With the assistance of AL, the role of the human oracle is either to label a datum or simply to correct the label from an automatic labeler. For the present work, we assume that correction is less costly than annotation from scratch; testing this assumption is the subject of future work. In AL, the learner leverages newly provided annotations to select more informative sentences which in turn can be used by the automatic labeler to provide more accurate annotations in future iterations. Ideally, this process yields accurate labels with less human effort.

Annotation cost is project dependent. For instance, annotators may be paid for the number of annotations they produce or by the hour. In the context of parse tree annotation, Hwa (2004) estimates cost using the number of constituents needing labeling and Osborne & Baldridge (2004) use a measure related to the number of possible parses. With few exceptions, previous work on AL has largely ignored the question of actual labeling *time*. One exception is (Ngai and Yarowsky, 2000) (discussed later) which compares the cost of manual rule writing with AL-based annotation for noun phrase chunking. In contrast, we focus on the performance of AL algorithms using different estimates of cost (including time) for part of speech (POS) tagging, although the results are applicable to AL for sequential labeling in general. We make the case for measuring cost in assessing AL methods by showing that the choice of a cost function significantly affects the choice of AL algorithm.

2 Benefit and Cost in Active Learning

Every annotation task begins with a set of unannotated items $\mathcal{U}$. The ordered set $\mathcal{A} \subseteq \mathcal{U}$ consists of all annotated data after annotation is complete or after available financial resources (or time) have been exhausted. We expand the goal of AL to produce the annotated set $\hat{\mathcal{A}}$ such that the benefit gained is maximized and cost is minimized.

In the case of POS tagging, tag accuracy is usu-

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 65–68,
Columbus, Ohio, USA, June 2008.

ally used as the measure of benefit. Several heuristic AL methods have been investigated for determining which data will provide the most information and hopefully the best accuracy. Perhaps the best known are Query by Committee (QBC) (Seung et al., 1992) and uncertainty sampling (or Query by Uncertainty, QBU) (Thrun and Moeller, 1992). Unfortunately, AL algorithms such as these ignore the cost term of the maximization problem and thus assume a uniform cost of annotating each item. In this case, the ordering of annotated data $\mathcal{A}$ will depend entirely on the algorithm's estimate of the expected benefit.

However, for AL in POS tagging, the cost term may not be uniform. If annotators are required to change only those automatically generated tags that are incorrect, and depending on how annotators are paid, the cost of tagging one sentence can depend greatly on what is known from sentences already annotated. Thus, in POS tagging both the benefit (increase in accuracy) and cost of annotating a sentence depend not only on properties of the sentence but also on the order in which the items are annotated.

Therefore, when evaluating the performance of an AL technique, cost should be taken into account. To illustrate this, consider some basic AL algorithms evaluated using several simple cost metrics. The results are presented against a random baseline which selects sentences at random; the learning curves represent the average of five runs starting from a random initial sentence. If annotators are paid by the sentence, Figure 1(a) presents a learning curve indicating that the AL policy that selects the longest sentence (LS) performs rather well. Figure 1(a) also shows that given this cost model, QBU and QBC are essentially tied, with QBU enjoying a slight advantage. This indicates that if annotators are paid by the sentence, QBU is the best solution, and LS is a reasonable alternative. Next, Figure 1(b) illustrates that the results differ substantially if annotators are paid by the word. In this case, using LS as an AL policy is worse than random selection. Furthermore, QBC outperforms QBU. Finally, Figure 1(c) shows what happens if annotators are paid by the number of word labels corrected. Notice that in this case, the random selector marginally outperforms the other techniques. This is because QBU, QBC, and LS tend to select data that require many corrections. Considered together, Figures 1(a)-Figure 1(c) show the significant impact of choosing a cost model on the relative performance of AL algorithms. This leads us to conclude that AL techniques should be evaluated and compared with respect to a specific cost function.

While not all of these cost functions are necessarily used in real-life annotation, each can be regarded as an important component of a cost model of payment by the hour. Since each of these functions depends on factors having a significant effect on the perceived performance of the various AL algorithms, it is important to combine them in a way that will accurately reflect the true performance of the selection algorithms.

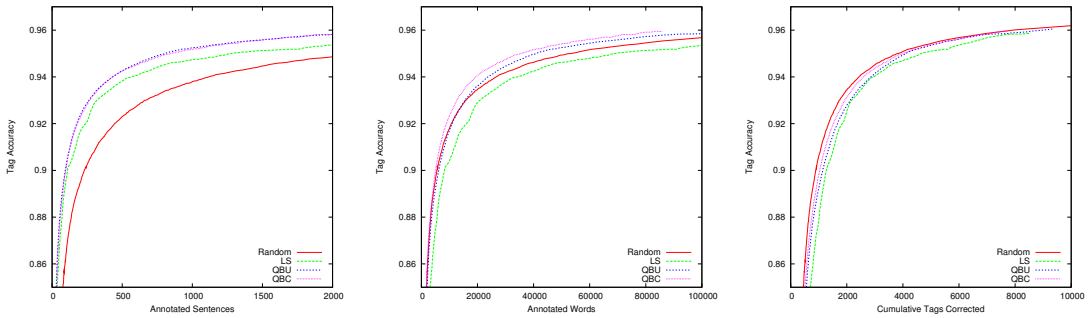
In prior work, we describe such a cost model for POS annotation on the basis of the time required for annotation (Ringger et al., 2008). We refer to this model as the "hourly cost model". This model is computed from data obtained from a user study involving a POS annotation task. In the study, timing information was gathered from many subjects who annotated both sentences and individual words. This study included tests in which words were prelabeled with a candidate labeling obtained from an automatic tagger (with a known error rate) as would occur in the context of AL. Linear regression on the study data yielded a model of POS annotation cost:

$$h = (3.795 \cdot l + 5.387 \cdot c + 12.57)/3600 \quad (1)$$

where h is the time in hours spent on the sentence, l is the number of tokens in the sentence, and c is the number of words in the sentence needing correction. For this model, the Relative Standard Error (RSE) is 89.5, and the adjusted correlation (R^2) is 0.181. This model reflects the abilities of the annotators in the study and may not be representative of annotators in other projects. However, the purpose of this paper is to create a framework for accounting for cost in AL algorithms. In contrast to the model presented by Ngai and Yarowsky (2000), which predicts monetary cost given time spent, this model estimates time spent from characteristics of a sentence.

3 Evaluation Methodology and Results

Our test data consists of English prose from the POS-tagged Wall Street Journal text in the Penn Treebank (PTB) version 3. We use sections 2-21 as



(a) (b) (c)

Figure 1: QBU, LS, QBC, and the random baseline plotted in terms of accuracy versus various cost functions: (a) number of sentences annotated; (b) number of words annotated; and (c) number of tags corrected.

initially unannotated data. We employ section 24 as the development test set on which tag accuracy is computed at the end of every iteration of AL.

For tagging, we employ an order two Maximum Entropy Markov Model (MEMM). For decoding, we found that a beam of size five sped up the decoder with almost no degradation in accuracy from Viterbi. The features used in this work are typical for modern MEMM POS tagging and are mostly based on work by Toutanova and Manning (2000).

In our implementation, QBU employs a single MEMM tagger. We approximate the entropy of the per-sentence tag sequences by summing over per-word entropy and have found that this approximation provides equivalent performance to the exact sequence entropy. We also consider another selection algorithm introduced in (Ringger et al., 2007) that eliminates the overhead of entropy computations altogether by estimating per-sentence uncertainty with $1 - P(\hat{\underline{t}})$, where $\hat{\underline{t}}$ is the Viterbi (best) tag sequence. We label this scheme QBUOMM (OMM = "One Minus Max").

Our implementation of QBC employs a committee of three MEMM taggers to balance computational cost and diversity, following Tomanek et al. (2007). Each committee member's training set is a random bootstrap sample of the available annotated data, but is otherwise as described above for QBU. We follow Engelson & Dagan (1996) in the implementation of vote entropy for sentence selection using these models.

When comparing the relative performance of AL algorithms, learning curves can be challenging to interpret. As curves proceed to the right, they can approach one another so closely that it may be difficult to see the advantage of one curve over another. For this reason, we introduce the "cost reduction curve". In such a curve, the accuracy is the independent variable. We then compute the percent reduction in cost (e.g., number of words or hours) over the cost of the random baseline for the same accuracy a:

$$redux(a) = (cost_{rnd}(a) - cost(a))/cost_{rnd}(a)$$

Consequently, the random baseline represents the trajectory $redux(a) = 0.0$. Algorithms less costly than the baseline appear above the baseline. For a specific accuracy value on a learning curve, the corresponding value of the cost on the random baseline is estimated by interpolation between neighboring points on the baseline. Using hourly cost, Figure 2 shows the cost reduction curves of several AL algorithms, including those already considered in the learning curves of Figure 1 (except LS). Restricting the discussion to the random baseline, QBC, and QBU: for low accuracies, random selection is the cheapest according to hourly cost; QBU begins to be cost-effective at around 91%; and QBC begins to outperform the baseline and QBU around 80%.

4 Normalized Methods

One approach to convert existing AL algorithms into cost-conscious algorithms is to normalize the results of the original algorithm by the estimated cost. It should be somewhat obvious that many selection algorithms are inherently length-biased for sequence labeling tasks. For instance, since QBU is the sum

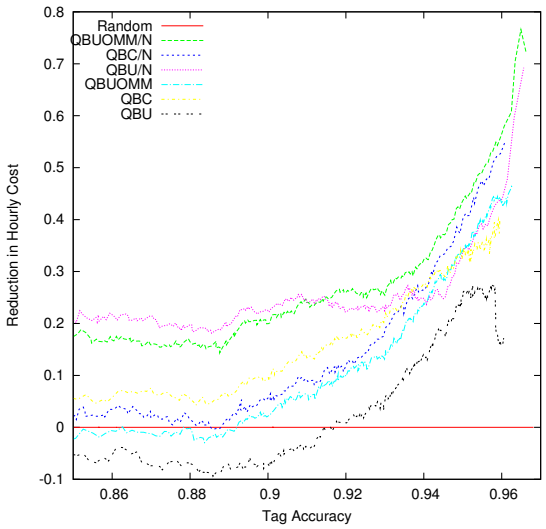


Figure 2: Cost reduction curves for QBU, QBC, QBUOMM, their normalized variants, and the random baseline on the basis of hourly cost

of entropy over all words, longer sentences will tend to have higher uncertainty. The easiest solution is to normalize by sentence length, as has been done previously (Engelson and Dagan, 1996; Tomanek et al., 2007). This of course assumes that annotators are paid by the word, which may or may not be true. Nevertheless, this approach can be justified by the hourly cost model. Replacing the number of words needing correction, c, with the product of l (the sentence length) and the accuracy p of the model, equation 1 can be re-written as the estimate:

$$\hat{h} = ((3.795 + 5.387p) \cdot l + 12.57)/3600$$

Within a single iteration of AL, p is constant, so the cost is approximately proportional to the length of the sentence. Figure 2 shows that normalized AL algorithms (suffixed with "/N") generally outperform the standard algorithms based on hourly cost (in contrast to the cost models used in Figures 1(a) - (c)). All algorithms shown have significant cost savings over the random baseline for accuracy levels above 92%. Furthermore, all algorithms except QBU depict trends of further increasing the advantage after 95%. According to the hourly cost model, QBUOMM/N has an advantage over all other algorithms for accuracies over 91%, achieving a significant 77% reduction in cost at 96.5% accuracy.

5 Conclusions

We have shown that annotation cost affects the assessment of AL algorithms used in POS annotation and advocate the use of a cost estimate that best estimates the true cost. For this reason, we employed an hourly cost model to evaluate AL algorithms for POS annotation. We have also introduced the cost reduction plot in order to assess the cost savings provided by AL. Furthermore, inspired by the notion of cost, we evaluated normalized variants of well-known AL algorithms and showed that these variants out-perform the standard versions with respect to the proposed hourly cost measure. In future work we will build better cost-conscious AL algorithms.

References

S. Engelson and I. Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *Proc. of ACL*, pages 319–326.

R. Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30:253–276.

G. Ngai and D. Yarowsky. 2000. Rule writing or annotation: cost-efficient resource usage for base noun phrase chunking. In *Proc. of ACL*, pages 117–125.

M. Osborne and J. Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proc. of HLT-NAACL*, pages 89–96.

E. Ringger, P. McClanahan, R. Haertel, G. Busby, M. Carmen, J. Carroll, K. Seppi, and D. Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proc. of Linguistic Annotation Workshop*, pages 101–108.

E. Ringger, M. Carmen, R. Haertel, K. Seppi, D. Lonsdale, P. McClanahan, J. Carroll, and N. Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proc. of LREC*.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proc. of CoLT*, pages 287–294.

S. Thrun and K. Moeller. 1992. Active exploration in dynamic environments. In *NIPS*, volume 4, pages 531–538.

K. Tomanek, J. Wermter, and U. Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. *Proc. of EMNLP-CoNLL*, pages 486–495.

K. Toutanova and C. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of EMNLP*, pages 63–70.