Assessing the Costs of Machine-Assisted Corpus Annotation Through a User Study

Eric Ringger, Marc Carmen, Robbie Haertel, Kevin Seppi, Deryle Lonsdale, Peter McClanahan, James Carroll, Noel Ellison

Computer Science Department
Brigham Young University
Provo, Utah 84602

E-mail: ringger@cs.byu.edu, marc.carmen@gmail.com, robbie_haertel@byu.edu, kseppi@byu.edu, lonz@byu.edu, petermcclanahan@gmail.com, jlcarroll@gmail.com

Abstract

Fixed, limited budgets often constrain the amount of expert annotation that can go into the construction of annotated corpora. Estimating the cost of annotation is the first step toward using annotation resources wisely. We present here a study of the cost of annotation. This study includes the participation of annotators at various skill levels and with varying backgrounds. Conducted over the web, the study consists of tests that simulate machine-assisted pre-annotation, requiring correction by the annotator rather than annotation from scratch. The study also includes tests representative of an annotation scenario involving Active Learning as it progresses from a naïve model to a knowledgeable model; in particular, annotators encounter pre-annotation of varying degrees of accuracy. The annotation interface lists tags considered likely by the annotation model in preference to other tags. We present the experimental parameters of the study and report both descriptive and inferential statistics on the results of the study. We conclude with a model for estimating the hourly cost of annotation for annotators of various skill levels. We also present models for two granularities of annotation: sentence at a time and word at a time.

1. Introduction

In the construction of annotated corpora, we are constrained by fixed budgets for expert annotation. Although a fully annotated corpus is required, we can afford only to label a subset. 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.

The current work focuses on part-of-speech tagging, although other annotation tasks can also benefit from the techniques discussed. In addition to a labeled corpus, we also aim to produce a probabilistic tagger that can accurately tag future texts. The annotation environment incorporates the probabilistic tagger in order to facilitate the annotation process: annotators are able to focus on correcting the tough cases missed by the automatic tagger while avoiding work on the cases tagged correctly in the pre-annotation. Future work will evaluate the comparative merits of annotation from scratch versus correction of pre-tagged text. In this work, a probabilistic part-of-speech tagger is incrementally trained from the labeled subset of a given corpus and employed to tag automatically the remainder of the corpus.

One important question that naturally arises in this setting of machine-assisted annotation is how to best focus the attention and expertise of the human annotator(s). One aspect of this question is: on which instances in the data should the annotators focus? This is the question addressed by active learning. Active Learning (AL) can be employed to reduce the costs of corpus annotation (Ringger et al., 2007; Tomanek, et al., 2007; Engelson & Dagan, 1996). Our previous work (Ringger et al., 2007) demonstrates that by applying active learning techniques, a state of the art tagging model can be trained on as little as one-half of the amount of data required by more traditional, less strategic annotation schemes to achieve the same levels of accuracy. With the assistance of AL, the role of the human oracle is either to label a datum of interest or simply to correct the label choice of an automatic labeler. AL directs an annotator's attention to those data which are likely to be maximally informative according to a given tagging model. In AL, the learner leverages newly provided annotations to select more informative sentences and to provide more accurate annotations in future iterations. Ideally, this process yields accurate labels with less human annotation. 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, Opper, and Sompolinsky, 1992) and Uncertainty Sampling (or Query by Uncertainty, QBU) (Thrun and Moeller, 1992).

A second aspect of the focus question is: at which granularity should the annotators direct their efforts? This paper focuses on this particular aspect and describes a user study designed specifically to assess the true cost of labeling a whole sentence or just a word at a time. Annotation cost is project-dependent. For instance, annotators may be paid by the hour or for the number of

annotations they produce (measured in words or sentences). With few exceptions, much of the previous work on AL has largely ignored the question of cost estimation. One exception is Ngai & Yarowsky (2000) who compare the cost of manual rule-writing with annotation using AL for noun phrase chunking.

In our previous work (Ringger et al., 2007), we assumed that the unit of annotation was a sentence. The assumption was based on a priori consideration of the nature of human input as an oracle for the POS tagging task. We reasoned that people gather contextual cues from a sentence in order to assemble the meaning of the whole. Consequently, we began with the assumption that a human annotator will usually require significant context from the sentence in order to label a single word with its POS label. We also reasoned that while focusing on a single word, the human may as well label (or correct the labels on) the entire sentence. The user study reported here questions the sentential assumption and tests its effectiveness against the word-at-a-time alternative.

In the following sections, we describe the experimental design of the study, the methods for data selection, the implementation of the web-based study itself, the user pool, and a statistical summary of the data produced by the study. The results allow us to make recommendations for the granularity of annotation and provide guidance for a model of the true hourly cost of annotation.

2. Experimental Design

2.1 Conditions

To address the question about the granularity of annotation, we construct an experiment designed to assess the cost of two alternatives: we ask users of a web-browser based interface to correct parts of speech on entire sentences and, separately, to correct the part of speech of individual words. For data annotation, time is money, so our measure of cost is the time required to annotate.

The reason for correction instead of de novo annotation is that in a machine-assisted annotation framework, we have access to labels from the statistical tagger. In this work we assume that machine assistance is always of some value, thus we do not test the case where words are tagged without assistance from the statistical tagger. Testing this assumption is the subject of future work.

For the second condition, correcting the tag on a single word, tags on neighboring words are hidden to avoid time wasted due to the distractions offered by tags on those neighboring words. We wish to avoid the potential cognitive load incurred on an annotator by the puzzlement related to seeing incorrect tags on words that cannot be corrected.

2.2 Control Variables

Our experiment includes two control variables. The first control variable is the accuracy of the tagger producing the tags to be corrected by the annotator. This variable allows us to determine the impact of tagger accuracy at various stages during the annotation process. In particular, for this first control variable, we employ statistical taggers of known error rates (created using a development corpus with known tags). Our study included tests using probabilistic taggers with accuracies: 50%, 75%, and 95%. This progression of increasing accuracy is typical in the process of active learning; hence, the data sheds some light on the time required to annotate in such circumstances. For tagging, we employed a probabilistic tagger, namely a Maximum Entropy Conditional Markov Model tagger (Ratnaparkhi, 1996; Toutanova& Manning, 2000; Toutanova et al., 2003). Such taggers are referred to alternatively in the literature as MaxEnt CMMs, MEMMs, or simply "MaxEnt" taggers.

The second control variable is the sentence length. Sentence length is discretized into three ranges: 1-15 words, 16- 21 words, and 22- 29 words. By selecting data for the study belonging to these three ranges, we are able to assess the impact of sentence length on the final cost.

2.3 Session Size

For the two conditions (entire sentences and word-a-a-time), three values of the first controlled variable (tagger accuracy), and three values of the second (discretized sentence length), we have eighteen different types of cases. Furthermore, we reasoned that each user should provide input on at least two examples of each of the eighteen case types. Consequently, in addition to practice and control sentences (4 each), each user responds to 36 cases, for a total of 44. As for the order of presentation, we decided to alternate between cases for each condition, beginning with a whole sentence annotation case, followed by a single word annotation case. The order of presentation was otherwise randomized without respect to the two control variables.

2.4 Data Selection

The study employs English prose from the Wall Street Journal portion of the Penn Treebank. This data set is well known, and the quality and shortcomings of its annotations are well understood.

Intuitively, data annotated by a more accurate model will require that fewer tags be corrected, thus requiring less annotation time. Similarly, longer sentences will be more costly than shorter sentences in general. As noted above, we determined that annotators could tag 36 cases, not including practice and control cases, in a reasonable amount of time. We estimated that we would be able to obtain data from a minimum of 25-30 annotators. We thus decided to produce 14 non-overlapping sets of 36 unique cases. From these, we produced 28 total "templates": each of the 14 sets are used in one template to be annotated word-at-a-time and in a second template as sentence-at-a-time. Consequently, the same cases are guaranteed to be annotated using both methods. Once

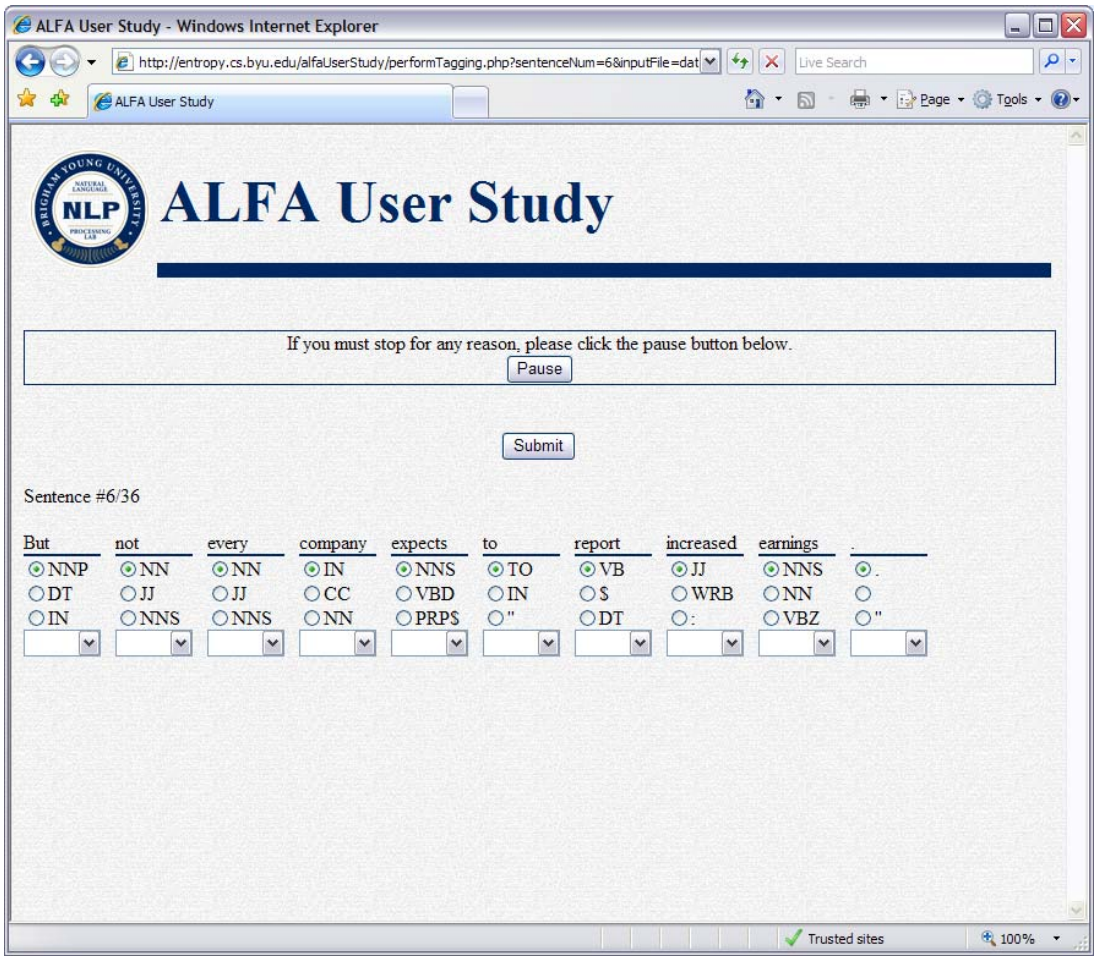

Figure 1. Web interface for the correction of tags for an entire sentence.

more than 28 users had participated in the study, templates were re-used.

We control for model accuracy by training three MaxEnt POS taggers of differing quality in an iterative fashion as follows. After the available training data (sections 2-21 of the Penn Treebank) was randomly ordered, a small batch of sentences was added to the (initially empty) annotated set, the model was retrained, and its accuracy on a held out set (section 24 of the PTB) was computed.

Model	Tag Accuracy	Unknown Word Tag Accuracy	Sentence Tag Accuracy
Tag50	54.3%	47.1%	0.4%
Tag75	75.1%	65.6%	1.5%
Tag95	95.3%	86.9%	36.9%

Table 1: Accuracy figures for the three models used to select sentences.

This process was repeated until the desired accuracy was achieved and the resulting model was saved. This was done for each of three desired accuracy levels: 50% (Tag50), 75% (Tag75), and 95% (Tag95); the actual accuracies obtained are shown in Table 1.

In order to control for length, we fit a log-normal distribution to the lengths of the sentences in the training data. The 25th, 50th, and 75th percentiles were used to create three distinct length buckets of [1, 15], [16, 21], and [22, 30] words, respectively. We excluded sentences having more than 30 words. We considered it better to present to the annotators a larger number of sentences in the time allotted to annotators than to have annotations on extremely long sentences.

The sentences in these length buckets are further divided into six equal parts in order to control for the other factors: two buckets (one for sentence at a time and one for word at a time annotation) for each of the three trained models (tag50, tag75 and tag90). The model with the appropriate accuracy was used to sort the sentences within each of the 18 buckets using the model corresponding to the bucket, and the first 28 sentences in each bucket were set aside. To ensure that a single template did not consist of all of the hardest sentences or words, we independently shuffled the

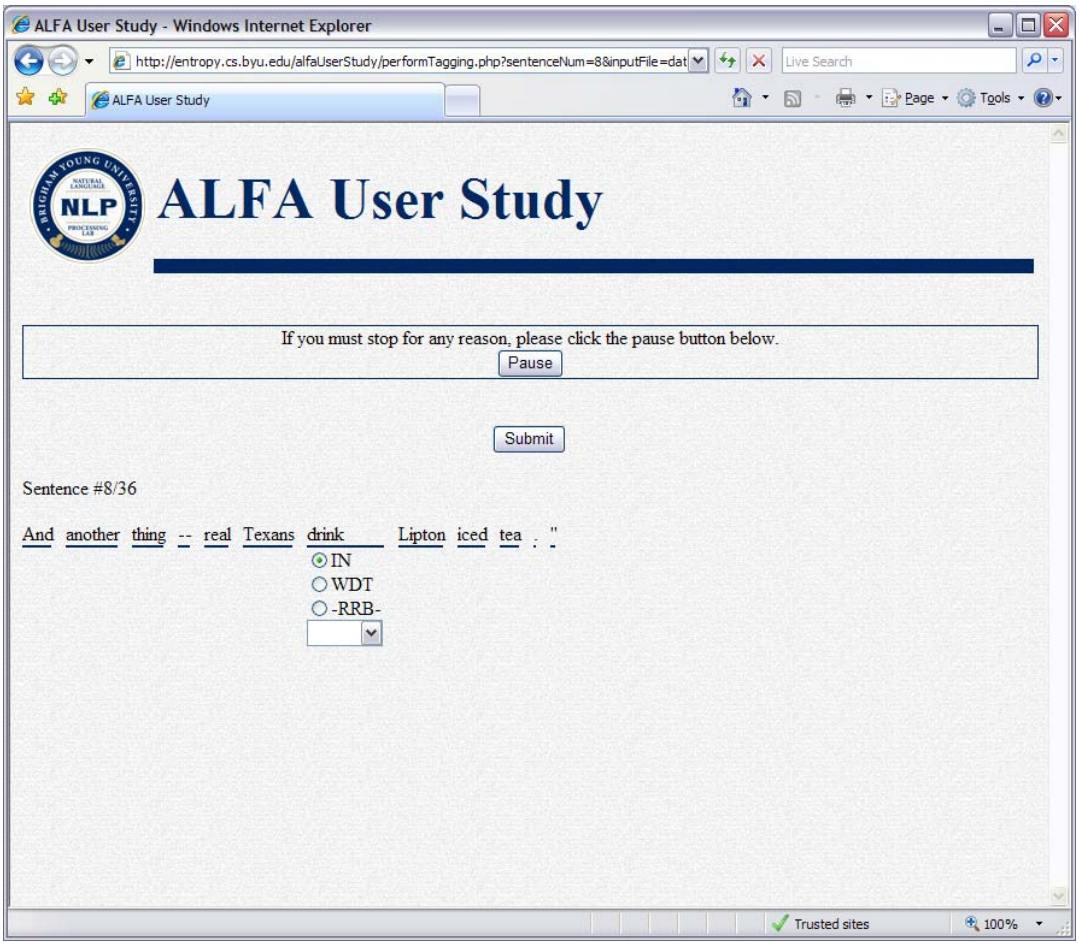

Figure 2. Web interface for the correction of tags for an individual word in sentential context.

28 remaining sentences in each of the buckets. A template was created by taking two sentences from each of the 18 buckets; half of them are presented the same way they were chosen (i.e., to be annotated as a full sentence if chosen using the sentence-at-a-time algorithm), and the other half were presented using the word-at-a-time method.

2.5 User Interface

The web study was implemented as a PHP application, and all session templates and results were encoded in XML. On the matter of how to present the cases to the user, we settled on the use of radio buttons for the top three tags with a drop-down for the other options. The top three tags are the three most likely correct tags as determined by the probabilistic tagger. The order of the tags in the drop-down list is static so that users do not have to "hunt" for the location of the desired tag. Figures 1 and 2 illustrate with screenshots of the interface for the annotation of sentences and words, respectively.

2.6 Subjects

The majority of the subjects for this study were undergraduate linguistics students in their third week of an undergraduate syntax course. They had received a general review of phrase types and word classes. The students received no special training in part of speech tagging, the tag set employed in the Penn Treebank, or even linguistic experimentation. The assignment was optional, and the equivalent value of a small homework assignment was awarded upon completion of the study. They were also rewarded for enrolling up to two additional people in the study. Ten of the 47 students were non-native speakers of English, and four of the 47 had participated in an earlier round of the study. When subjects finished the study they were given a brief survey in which they were asked several questions related to their ability and their performance in the study, as described in greater detail below.

3. Descriptive Statistics

We present the results of this study in two parts: first we introduce the data gathered and present descriptive statistics. Second, we model the data and draw conclusions using inferential statistics.

The web-based annotation tool described in the preceding sections gathered the following data during the annotation study:

Length: the number of tokens in the sentence; when annotating a single word it is the length of the sentence in which the word appears

Time: the time in seconds that the subject spent on the current case

Subject Accuracy: the percentage of tokens correctly tagged by the subject. When annotating a single word this is either 0% or 100%

Location: Index of the current case in the session

Tagger Accuracy: The percentage of words correctly tagged by the automatic tagger in the sentence. When annotating a single word this is either 0% or 100%

Number Needing Correction: the number of words in the case needing correction

Percent Done: percentage of the cases assigned to the current subject already encountered

Conditional Entropy:

- for whole sentence annotation, an estimate of the total tag sequence entropy given the words in the current sentence
- for single word annotation, the entropy of the tag distribution for the current word

From Tagger: the accuracy (50, 75, or 95) of the tagger providing the candidate tags

Native English Speaker: a 0/1 indicator of whether the subject is a native English speaker

Previously Participated in Study: a 0/1 indicator of whether the subject was part of a previous (similar) tagging exercise

Self Evaluation Tagging Proficiency: a 1/2/3/4/5 indicator of the subject self-evaluation of tagging proficiency

Self Evaluation of Performance in Study: a 1/2/3/4/5 indicator of the subject self-evaluation of tagging accuracy in this study

Descriptive statistics on these attributes are shown in Table 2. The final row in the table measures the time required per tag. For the word-at-a-time case, this is just the value of the Time variable above. For the sentence-at-a-time case, the time per tag is the ratio of the time per sentence by the length of the sentence.

4. Hourly Cost Models

We are interested primarily in linear models predictive of the time required for annotation/correction tasks. Based on the data from the annotation user study, we derive such models. We will refer to these models as "hourly cost models". First we focus only on the data annotated/corrected one sentence at a time. There were 1046 annotated sentences in the data from the user study. We discarded extreme outliers having time less than or equal to five seconds or greater than or equal to 1000 seconds. Such outliers are probably best explained by the failure of a subject to use the "pause" button in the web interface or by negligent speeding through the study. This left 906 sentences. The hourly cost model computed on that data by means of linear regression is as follows:

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

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. The model uses only a small subset of the raw statistics available in the annotation study for two reasons: first, some variables (e.g., proficiency assessment) are not included because we explicitly wish to assume that tagging will be conducted by a mix of people with tagging skills similar to the mix of skills tested in the user study. Second, some variables fail to have a statistically meaningful effect on the resultant model. We employed the Bayesian Information Criterion (as implemented in the LEAPS package in R) to assess which of the variables listed in Section 3 should be included in the model. For this model, the Residual Standard Error (RSE) is 89.5, and the adjusted correlation (R^2) is 0.181.

The model has an intuitive interpretation: the annotator considers each word and decides whether or not it needs to be corrected (3.795 seconds per word); only words needing correction are changed (5.387 seconds per correction). Additionally, there is 12.57 seconds of overhead per sentence. In contrast to the model presented in Ngai and Yarowsky (2000) which predicts monetary cost given time spent, this model estimates time spent from characteristics of a sentence. Many of the costs employed in other work (Hwa, 2000; Osborne & Baldridge, 2004) can be seen as estimating only some portion of the hourly cost. These distinctions make our work a novel contribution.

This model reflects the abilities of the annotators in the study may not be representative of expert annotators hired for other annotation work. A better model, linear or otherwise, based on data from other annotators could be employed, but the methodology employed in this study could be applied directly.

We also found the following relationships. Interestingly, participation by non-native speakers of English did not appear to affect accuracy but does affect completion time. As noted above, each subject was asked to rate his proficiency in tagging; self evaluation is statistically significant in relationship to the subject's accuracy. Also, a subject's self evaluation and correction accuracy are correlated. Whether or not a subject had participated in a similar previous experiment had no statistically significant impact on subject accuracy or on time to completion. "Conditional entropy" of the tag sequence distribution given the words has a negative effect on the subject's accuracy: as entropy increases, subject accuracy decreases.

We have also entertained other questions regarding the sentence-at-a-time data. If we consider only the self-rated experts in the data set, in other words if we examine only the data from subjects whose proficiency rating is 3, then we are limited to 300 sentences in the data. We apply the

Label	Sentence					Word				
	Mean	Std Dev	Median	Min	Max	Mean	Std Dev	Median	Min	Max
Length	19.86	5.74	20	4	30	19.85	5.77	20	4	30
Time (Seconds)	137.02	141.69	97.69	0.02	1448.67	19.4	21.92	14.09	0.14	299.23
Subject Accuracy	78.37	16.18	80.95	5	100	69.92	45.88	100	0	100
Location	22.91	12.66	23	1	50	22.42	12.93	22	2	49
Tagger Accuracy	63.68	22.86	65	5	100	32.21	46.75	0	0	100
Number Needing Correction	7.21	5.07	7	0	23	7.35	5.04	7	0	23
Percent Done	51.7	28.54	52.27	2	100	50.63	29.17	50	4	98
Conditional Entropy	30.51	16.81	31.09	3.74	75.23	30.64	16.83	31.09	3.74	75.23
From Tagger	73.74	16.68	75	50	95	73.62	16.66	75	50	95
Native English Speaker	0.79	0.41	1	0	1	0.79	0.41	1	0	1
Previously Participated in Study	0.09	0.29	0	0	1	0.09	0.28	0	0	1
Self Evaluation Tagging Proficiency	2.11	0.72	2	1	3	2.1	0.72	2	1	3
Self Evaluation of Performance in Study	2.77	0.83	3	1	5	2.76	0.83	3	1	5
Time per Tag	6.98	7.01	5.09	0	68.98	19.4	21.92	14.09	0.14	299.23

Table 2: Statistics for Sentence-at-a-time and word-at-a-time annotation

same linear regression techniques, and the resulting hourly model is:

$$h = (4.261 \cdot l + 4.683 \cdot c - 5.579)/3600$$

As before, l = length, and c = the number of tokens needing correction. For this model, the RSE is 76.57, and the adjusted R^2 is 0.2345.

If we consider only the beginners (self rating is 2 or lower), then we are limited to 606 sentences in the data. The resulting model is:

$$h = (3.441 \cdot l + 3.441 \cdot c + 20.752)/3600$$

For this model, the RSE is 94.98, and the adjusted R^2 is 0.1622.

Next, we analyze the sentences having high annotation accuracy. The number of sentences having an annotation accuracy of at least 95% (annotations come from the user study subjects) is 111. The linear model on this subset of the data is:

$$h = (0.711 \cdot l + 5.174 \cdot c + 47.876)/3600$$

For this model, the RSE is 76.64, and the adjusted R^2 is 0.1109.

For the word-at-a-time data, there were 1035 cases of annotated words. We discarded extreme outliers having time less than or equal to one second or greater than or equal to 200 seconds. This left 915 sentences. The hourly cost model computed on the word-at-a-time data by means of linear regression is as follows:

$$h = (14.193 + 5.670 \cdot b)/3600$$

where b is a binary indicator reflecting whether or not the word in question actually needed correction. As before, this model was preferred by the Bayesian Information Criterion. For this model, the RSE was 15.76, and the adjusted R^2 is 0.0256. The next best model according to BIC included the length of the sentence in which the word occurred, as in the sentence-at-a-time model. Intuitively, length may play a small role, for instance, affecting the time to scan for and find the word to be tagged; also, in longer sentences larger context may be needed by the subject to disambiguate more distant co-reference.

5. Future Work

Based on this analysis, we have simple linear models with which to predict the time required to annotate data using each presentation technique, whether sentence-at-a-time or word-at-a-time. Our future work focuses on the application of these results in the context of active learning. Our previous work demonstrates that by applying active learning techniques, a state of the art tagging model can be trained on as little as one-half of the amount of data required by more traditional machine-assisted annotation schemes to achieve the same levels of accuracy. These results assumed that cost was measured in terms of the number of sentences annotated. We will also evaluate the cost of annotation with the new models and assess overall cost reductions in terms of time

and therefore money. It is ultimately expenditures of money that are limited by our project budgets.

We also plan to apply the models of annotation cost derived here in the current work to select the presentation style for annotation. By adaptively presenting cases one word-at-a-time or one sentence-at-a-time, we expect to be able to further minimize annotation time and, therefore, cost.

References

Engelson, S. and Dagan, I. (1996). "Minimizing manual annotation cost in supervised training from corpora." ACL. Pp. 319-326.

Hwa, R. (2000). "Sample selection for statistical grammar induction". In Proceedings of EMNLP/VLC-2000. Pp. 45-52.

Lewis, D., and Catlett, J. (1994). "Heterogeneous uncertainty sampling for supervised learning." ICML.

Lewis, D., and Gale, W. (1995). "A sequential algorithm for training text classifiers: Corrigendum and additional data." SIGIR Forum, 29 (2), Pp. 13-19.

Ngai, G. and Yarowsky, D. (2000) "Rule Writing or An-notation: Cost-efficient Resource Usage for Base Noun Phrase Chunking." ACL. Pp. 117-125.

Osborne, M., & Baldridge, J. (2004). "Ensemble-based AL for Parse Selection". HLT-NAACL 2004, Boston, Massachusetts, USA. Pp. 89—96.

Ratnaparkhi, A. (1996). "A Maximum Entropy Model for Part-Of-Speech Tagging." EMNLP.

Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., Seppi, K., & Lonsdale, D. (2007). "Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation". ACL Linguistic Annotation Workshop (LAW) 2007. Prague, Czech Republic.

Seung, H., Opper, M., & Sompolinsky, H. (1992). "Query by committee". COLT. Pp. 287-294.

Thrun S., and Moeller, K. (1992). "Active exploration in dynamic environments." NIPS.

Tomanek, K., Wermter, J., & Hahn, U. (2007). "An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of An-notated Data." EMNLP. Pp. 486-495.

Toutanova, K. & Manning, C. (2000). "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger." EMNLP. Pp. 63-70.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." HLT-NAACL. Pp. 252-259.