Journal of the American Medical Informatics Association, 26(11), 2019, 1314–1322
doi: 10.1093/jamia/ocz102
Advance Access Publication Date: 11 July 2019
Research and Applications

Research and Applications

Cost-aware active learning for named entity recognition in clinical text

Qiang Wei,[1] Yukun Chen,[2] Mandana Salimi,[1] Joshua C Denny,[3,4] Qiaozhu Mei,[5] Thomas A. Lasko,[3] Qingxia Chen,[3,6] Stephen Wu,[1] Amy Franklin,[1] Trevor Cohen,[7] and Hua Xu[1]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA, [2]Pieces Technologies Inc, Dallas, Texas, USA, [3]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA, [4]Department of Medicine, Vanderbilt University, Nashville, Tennessee, USA, [5]School of Information, University of Michigan, Ann Arbor, Michigan, USA, [6]Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, USA, and [7]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington, USA

Corresponding Author: Trevor Cohen, MBChB, PhD, Department of Biomedical Informatics and Medical Education, University of Washington, 850 Republican Street, Seattle, WA 98109, USA; cohenta@uw.edu and Hua Xu, PhD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 870, Houston, TX 77030, USA; hua.xu@uth.tmc.edu

Received 28 November 2018; Revised 17 May 2019; Editorial Decision 22 May 2019; Accepted 5 June 2019

ABSTRACT

Objective: Active Learning (AL) attempts to reduce annotation cost (ie, time) by selecting the most informative examples for annotation. Most approaches tacitly (and unrealistically) assume that the cost for annotating each sample is identical. This study introduces a cost-aware AL method, which simultaneously models both the annotation cost and the informativeness of the samples and evaluates both via simulation and user studies.

Materials and Methods: We designed a novel, cost-aware AL algorithm (Cost-CAUSE) for annotating clinical named entities; we first utilized lexical and syntactic features to estimate annotation cost, then we incorporated this cost measure into an existing AL algorithm. Using the 2010 i2b2/VA data set, we then conducted a simulation study comparing Cost-CAUSE with noncost-aware AL methods, and a user study comparing Cost-CAUSE with passive learning.

Results: Our cost model fit empirical annotation data well, and Cost-CAUSE increased the simulation area under the learning curve (ALC) scores by up to 5.6% and 4.9%, compared with random sampling and alternate AL methods. Moreover, in a user annotation task, Cost-CAUSE outperformed passive learning on the ALC score and reduced annotation time by 20.5%–30.2%.

Discussion: Although AL has proven effective in simulations, our user study shows that a real-world environment is far more complex. Other factors have a noticeable effect on the AL method, such as the annotation accuracy of users, the tiredness of users, and even the physical and mental condition of users.

Conclusion: Cost-CAUSE saves significant annotation cost compared to random sampling.

Key words: natural language processing, active learning, electronic health records, named entity recognition, user study

INTRODUCTION

Supervised machine learning (ML) models have achieved state-of-the-art performance across a range of clinical natural language processing (NLP) tasks,[1,2] but statistical NLP systems often require large numbers of annotated samples in order to build high performance ML models. Constructing large-scale, high-quality corpora is time consuming and costly, particularly in the medical domain where corpus-building often requires manual annotation by domain experts. Therefore, methods that can help build high-performance ML models but require fewer annotations are highly desirable in clinical NLP.

Active learning (AL) systems attempt to prioritize more informative samples for annotation during an iterative training process; this contrasts with the standard passive learning (PL) strategies, such as random sampling. AL approaches may select samples based on diversity (samples that are least similar to already-annotated ones),[3,4] uncertainty (the samples the model considers most difficult to categorize may be most informative),[5] query-by-committee (samples that multiple systems are least in agreement on),[6] or other approaches.[7]

The contribution of the current work is twofold. First, our AL approach is to consider annotation costs while selecting AL samples, extending CAUSE (Clustering And Uncertainty Sampling Engine)[8] to weigh these costs against informativeness in a new model we call Cost-CAUSE. Second, in addition to a typical simulated AL evaluation where we measure the area under the learning curve (ALC), we also complete a user study of real annotators using our AL system.

Active learning successes without cost modeling

AL has been widely studied in the biomedical context and in the general domain. Settles and Craven conducted a large-scale general-domain evaluation of multiple AL methods with a number of approaches giving relatively robust performance.[9] Chen et al demonstrated that AL outperformed random sampling for a simulated clinical named entity recognition (NER) task,[4] while Kholghi et al showed that AL in the clinical domain could reach the same ML accuracy using only 54% (i2b2/VA 2010) and 76% (ShARe/CLEF 2013) of the total number of concepts in the training data.[10] Chen et al clustered sentences into groups based on their content, then combined the uncertainty and diversity of samples to query samples from an unlabeled pool and showed a better performance compared with traditional AL algorithm.[8]

Despite the fact that these studies demonstrated the potential of AL, they were conducted in simulated environments, and assumed that annotation costs for each sample were identical. They maximized informativeness (ie, how much a sample will contribute to learning an ML model), divided by the number of sentences or words. However, in reality, annotation cost (ie, the time required by an annotator) can be very different from 1 sample to another and from 1 user to another user. So, the number of annotated examples is a surrogate estimate of cost and may not accurately reflect the actual time required for annotation—which is the primary concern for practical purposes.

Active learning and real-world annotation time

Addressing this issue of real-world annotation time, Settles et al[11] collected several corpora along with sample-level annotation times to evaluate real-world AL performance, and found that observed costs are highly variable across instances. Chen et al[8] developed an AL annotation system to sample sentences for users, concluding on the basis of user studies that cost-agnostic AL approaches may perform no better than random sampling on the measurement of annotation time, but improved learning curves are achievable if the cost variables can be appropriately taken into account. Kholghi et al[12] recruited 4 users to compare various AL methods with random sampling. The AL methods in their tests reduced annotation time by 28% compared with random sampling.

However, while these studies did advance understanding of the effects of AL on annotation time, the AL approaches evaluated did not explicitly address annotation time in their querying strategies. Furthermore, the experiments involved a relatively small number of users, did not compare AL with PL, and did not control as extensively for noise factors.

Cost-conscious active learning

Introducing a cost (ie, time) variable into an AL strategy is 1 way to reflect real-world annotation time. A number of studies proposed methods to model the cost of a sample and balance that cost against informativeness. Haertel et al[13] presented a practical cost-conscious AL approach motivated by the business concept of return on investment, and showed a 73% reduction in hourly cost as compared with random sampling on a part of speech (POS) tagging task. Tomanek et al[14] summarized and compared several methods that incorporated cost variables into AL. Both studies found that using the ratio between informativeness and cost was an effective way to incorporate a cost variable into AL.

However, the estimation of time-specific cost variables for AL is not a trivial task, since AL models do not know the cost of a sentence before it has been annotated by a user in practice. Consequently, a predictive model for annotation time is required. In the open domain, time-specific annotation cost models have been proposed for NLP tasks like POS tagging,[15] text classification,[16] and NER.[17] All of these studies took count characteristics of samples into consideration including number of words and sentences at the document level. Besides those variables, Arora et al[16] also used variables describing annotator characteristics, and Tomanek et al[17] incorporated the semantic and syntactic complexity of sentences into a cost model. Haertel et al contributed a theoretical analysis of cost model-based AL and presented a simulated study for POS tagging, which showed it can indeed successfully reduce total annotation time and be considered a viable option for machine-assisted annotation.[18] However, though these previous studies did utilize cost-conscious AL methods, they did not evaluate these methods in the context of user studies. Consequently, these studies do not address important aspects of the complexity of annotation in production environments, such as differences in annotation quality across users and the element of user fatigue—both of which affect the estimated performance of AL.

MATERIALS AND METHODS

Data set

In this study, we used data from the 2010 i2b2/VA challenge, preserving the original training and test splits of 349 clinical documents (20 423 unique sentences) and 477 clinical documents (29 789 unique sentences).[19] There were 3 types of medical entities annotated in each sentence: problem, treatment, and test (see details in the Supplementary Table 1). We also utilized the annotation time for 2 users (887 sentences for user A, and 891 for user B) obtained from a previous study.[8] These annotation times were used to

Table 1. Features in the cost model. Named entities are in boldface

Sentence	***MRI*** by report showed ***bilateral rotator cuff repairs*** and he was admitted for ***repair of the left rotator cuff***.				
Categories	Count			Lexicon	Syntactic
Feature	Number of words	Number of entities	Number of entity words	Inverse Document Freq.	Entropy of POS tag
Value	20	3	11	35.36	2.28

develop our cost model. The training set was used in the simulation study and for training annotators in the user study. The test set was used in the user study.

Cost-CAUSE

We propose Cost-CAUSE as an approach to identify more informative, less costly sentences. While we follow the CAUSE[8] query strategy to select unlabeled sentences for annotation, we score sentences using the ratio *Informativeness(s)/Cost(s)* between the informativeness of a sentence s and its estimated annotation time, similar to other cost-conscious approaches.[7,9] Cost-CAUSE is encapsulated by the following pseudocode in Box 1. The ranked sentence set S maintains a balanced distribution across topics while selecting high informativeness per cost (IPC) samples, and top sentences in S will be used for annotating.

Box 1: Cost-CAUSE algorithm.

1. Cluster sentences s into groups g according to their topics.
2. Calculate IPC for each sentence:

$$IPC(s) = \frac{Informativeness(s)}{Cost(s)}$$

3. Calculate averaged IPC for each group g_i:

$$Avg\ IPC(g_i) = \frac{\sum_{s \in g_i} IPC(s)}{\#\{sentences\ in\ g_i\}}$$

4. Ranked group list $\rightarrow G$: $g_i,\ g_2,\ \ldots,\ g_n$
5. For g_i in G:
   select sentence s with highest IPC in g_i put s into ranked sentence set S remove s from g_i

While *Informativeness(s)* is determined by established querying algorithms (eg, entropy of words in the sentence, entropy of entities in the sentence, least confidence of the sentence[4]), our *Cost(s)* model is unique and we thus describe it in further detail below.

Annotation cost model

We developed a cost model to estimate annotation cost for 1 sentence based on features selected to capture the basic characteristics, lexical complexity, and syntactic complexity of the sentence. Motivated by psycholinguistic literature showing that linear combinations of estimates can model the time it takes to read words,[20,21] the linear model for annotation time estimation is given by the following formula:

$$Cost(s) = c_0 + \sum_i c_i f_i(s)$$

where $f_i(s)$ is the value of feature i for sentence s, and coefficients c_i are parameters learned during training.

Table 1 shows all the features used in the study. The *Count* features reflect the characteristics of sentences, such as their length and the number of entities and entity words they contain. To model syntactic complexity, we developed a *Syntactic* feature based on the probability of POS bigrams. The underlying assumption is that relatively infrequent POS sequences may take longer to process. Specifically, the *POS Tag Entropy* feature $H(s)$ is based on the corpus-derived probabilities of POS bigrams. $P(s)$ represents the probability of a sentence s, which is estimated as the product of the probabilities of POS tags of each bigram in s. Formally, $H(s)$ is calculated from this as follows, where n is the length of the sentence s and p_i is the POS tag of the ith word in the sentence s:

$$H(s) = -\frac{1}{n} log\ P(s) = -\frac{1}{n} \log_e \left(\sum_{p_i \in s} P(p_i | p_{i-1}) \right)$$
$$= -\frac{1}{n} \log_e \left(\sum_{p_i \in s} \frac{P(p_i, p_{i-1})}{P(p_{i-1})} \right)$$

Lexical features model how difficult it is for annotator to understand each word in a sentence at the level of meaning. The cumulative inverse document frequency (IDF, with sentences as "documents") is used to measure the lexical complexity on the assumption that infrequently encountered terms may take longer to process.

We calculated the value of each feature for 887 and 891 sentences respectively and used them to estimate the c_i values, fitting linear models to the annotation times of these 2 users in the simulation.

Simulation

A simulation was performed in order to evaluate the Cost-CAUSE model. For AL models in this simulation, an evaluation set E was set aside for evaluating an incrementally trained ML model; the remaining data were initially considered unlabeled data U. Some samples from U were subsequently selected (eg, via our Cost-CAUSE scoring) and "labeled" with the gold standard label, moving these samples to a set of labeled data L. With the labeled data, a ML model was trained and evaluated on E. More samples from U were selected and "labeled" and trained on until an artificial time limit of 2 simulated hours was reached. We used the cost model to estimate annotation time for each sentence in the labeled set L and summed these estimates, yielding our simulated time.

In our tests, we randomly split the 2010 i2b2/VA NLP challenge training data set with 20 423 unique sentences into 5 folds for cross-validation. Out of 5 folds, 4 were used as the initial unlabeled pool U, while the remaining fold was the independent test set for evaluation E. The workflow of the simulation is detailed as follows in Box 2.

The same procedure was used for all tested AL systems; of course, a PL random sampling method did not make use of the Cost-CAUSE derived labeled/unlabeled data partitions. The learning

Box 2: The workflow of the simulation.

1. Initialization:
   a. Fit a linear cost model *Cost*(*s*) to each user.
   b. Split the data into 5 folds.
   c. Put 4 folds of sentences into unlabeled pool *U*. The remaining fold is the evaluation set *E*.
2. Loop until estimated annotation time reaches 120 minutes:
   a. Rank sentences in the unlabeled pool *U* (eg using Cost-CAUSE)
   b. Select the top 5 sentences from ranked sentences and put them into labeled pool.
   c. Use sentences as input to train a CRF model.
   d. Use model from step 2c to predict sentences in the unlabeled pool *U*.
   e. Calculate the performance of the model from step c using evaluation set.

curves that plot F-measures vs. estimated annotation time were generated to visualize the performance of different methods. The area under the learning curve (ALC; see equation in the Supplementary Material) was used to compare the performance of AL or PL methods. The ALC measured the expected F-measure that the ML model (trained by samples selected by AL or PL methods) can achieve, within a given time window.

User study

To further validate the utility of the Cost-CAUSE method, we conducted a user study to compare overall NER annotation times using Cost-CAUSE vs random sampling. This study received the Institutional Review Board approval from Committee for the Protection of Human Subjects at The University of Texas Health Science Center at Houston (HSC-SBMI-14-0678). The participants were recruited in the University of Texas Health Science Center at Houston, and met the following conditions: (1) they were medical or nursing students; (2) they had experience working with clinical notes written in English. The actual training and evaluation were conducted in 3 phases, designed for consistency in annotation skill level and environment.

1. **Phase I**. 20 participants took basic NER annotation training, and were tested to characterize their level of accuracy.
2. **Phase II**. All 20 participants entered Phase II, but 8 participants discontinued the user study for personal reasons. Data used in Phase II came from the i2b2/VA. Participants took a further 90 minutes of training: additional learning materials, discussion of guidelines, and further annotation cases. Then all participants completed 3 sessions of practice, annotating in our system for 15 minutes and reviewing their annotations. Finally, they took a 1-hour test to determine whether they were eligible for Phase III, in which they annotated an unseen set of sentences. We chose 10 participants for the next phase, with a minimum annotation quality (measured by F1 score) of 0.67. Also, data from the test were used for fit parameters of a cost model for each participant.
3. **Phase III**. Phase III was conducted in 2 days. For each day, participants reviewed their annotation from the Phase II test for a half hour to warm up. Then, they took the annotation test using our annotation system with the i2b2 2010 data for 120 minutes. The test was in 3 40-minute sessions and there was a 15-minute break between any 2 sessions. A total of 10 medical experts (nurses, medical students, and physicians) completed Phase III, but 1 participant was subsequently removed from further analysis due to lower annotation quality. No users have unique combinations of background and medical training (Supplementary Table 2). Data used in Phase III came from the original test set of the i2b2/VA, and the data set for evaluating models trained from users' annotation in this phase came from the original training set of the i2b2/VA.

Annotation system design to control potential bias. User studies comparing active and passive learning are inherently vulnerable to bias on account of differences in annotation times across users and differences in annotation times for the same users, which may increase as they gain experience or decrease as they become fatigued during the annotation. In order to mitigate for these biases when comparing AL to PL, we designed a system (see details in Supplementary Material) that can use either AL or PL to query sentences for users to annotate, while adjusting its selection among these methods to ensure that (1) users spend similar time on sentences from AL and PL, and (2) the sentences are presented to users at the same point in a session (mitigating for fluctuations in user annotation speed over time).

RESULTS

Simulation studies

We trained the baseline and the proposed annotation cost models for 2 users using their own annotated data (Table 2). We started with number of words in a sentence (the NOW model) as a baseline, which fit poorly to both user A and B. Incorporating both the number of entities and the number of entity words in the cost model improves the R^2 value (the COUNT model), and adding IDF (COUNT_LEXICAL) and other syntactic complexity features (COUNT_LEXICAL_SYNTACTIC) further improved the model's fit.

In Figure 1 (learning curves) and Table 2 (ALC scores), the Cost-CAUSE method shows better ALC than that of RADOM, Uncertainty, and CAUSE. Among cost-aware algorithms, the model with COUNT_LEXICAL_SYNTACTIC achieved the best ALC score, with improvements of 5.6% for both users for random sampling and 3.3% for user A and 4.9% for user B for CAUSE. Comparing the cost-agnostic AL methods, we note that these perform no better than random sampling when evaluated using annotation time (rather than number of words or sentences).

We also observed different characteristics of the sentences queried by different methods (Table 4 in Supplementary Material). Each technique (RANDOM, Uncertainty, CAUSE, and Cost-CAUSE) varies in its average sentence length (11, ~42, 27, and 11–13), entities per sentence (1.25, 6.7, 3.9, and 13), and entity density (0.24, 0.37, 0.36, and 0.32–0.34). Uncertainty and CAUSE seem to select longer sentences, which explains why their performance is overestimated when only number of sentences is considered; while Cost-CAUSE selects sentences with similar length as RANDOM.

User study of Cost-CAUSE

Performance of AL and PL. AL (Cost-CAUSE) outperformed PL (random selection) for 8 of 9 users in ALC scores (Table 3 and Figure 2), and the score for AL was significantly larger than that for PL (Wilcoxon signed-rank test, $P < .01$). At 120 minutes, the ML model

Table 2. R-squared value for cost models and the area under the learning curve (ALC) for each method – Passive (PL) vs Active (AL), cost-agnostic vs cost-aware – in a 2-hour simulation study. Best results for each user are in boldface

	R^2		ALC	
	User A	User B	User A	User B
PL: RANDOM	–	–	0.621	0.620
AL: Uncertainty (LC)	–	–	0.595	0.590
AL: CAUSE	–	–	0.635	0.624
Cost-CAUSE variants				
AL: NOW[a] (baseline)	0.668	0.459	–	–
AL: COUNT[a]	0.791	0.530	0.650	0.650
AL: COUNT_LEXICAL[a]	0.818	0.543	0.652	0.649
AL: COUNT_LEXICAL_SYNTACTIC[a]	**0.823**	**0.545**	**0.656**	**0.654**

Abbreviations: AL, active learning; ALC, under the learning curve; LC, least confidence; PL, passive learning.

[a]Results of 4 Cost-CAUSE variants were shown, which used cost models including different features to estimate annotation time. The text after colon shows the features in the cost model.

trained on AL sentences had a better performance than that from PL sentences in 7 of 9 users. To test whether AL is significantly different from PL in terms of performance of the ML model as captured by the learning curves, we performed a Wilcoxon signed-rank test for each user, and AL significantly outperformed PL in terms of ALC scores for 6 of 9 users ($P < 10^{-3}$).

Annotation performance. Table 4 shows the characteristics of the annotation processes using AL (Cost-CAUSE) and PL (random sampling). The annotation quality F-measure was estimated by comparing user annotations to the reference standard. While users maintained at least 0.70 F-measure on annotation quality, there was an observable difference between AL and PL (0.748 for AL and 0.798 for PL on average; a median of 0.74 for AL and 0.79 for PL). Annotation qualities of 3 users for AL sentences were much lower than PL (user 1, 3, and 4, ~0.08 lower). Users spent a longer time annotating words in AL sentences (33.01–73.07 vs 40.47–92.22 words/minute) and annotated fewer AL sentences within 120 minutes. AL sentences were slightly longer (12.44 vs 11.38 words/sentence on average), contained more entities (2.14 vs 1.39 entities/sentences on average), had a higher entity density (0.34 vs 0.26 on average), and thus were perhaps more difficult for users.

There was a decrease in annotation quality for AL from 40 minutes (Figure 2, blue dashed lines), where annotation quality for AL clearly falls below PL. This may be because sentences selected by AL become progressively harder to annotate as sentences must be more atypical to qualify as "informative" as the model evolves (further exploration is in the Supplementary Material).

Annotation effort saved by Cost-CAUSE. Consider a complementary measurement of users' annotation effort: how much annotation is necessary (in minutes, number of sentences, and number of words) to reach a target performance F-measure of 0.67? For users 1 and 4, the AL model took more time to reach target performance than PL. For another 2 users, the PL model never reached the target performance at the end of 120 minutes, while the AL model did. For the remaining 5 users, AL reduced the annotation time to reach the target performance by 20.5%—30.2% and reduced the number of sentences and words annotated at target performance by 43%–49.4% and 37.6%–44.4%, respectively (Supplementary Table 5).

Interestingly, although Cost-CAUSE did not reduce annotation time for User 1 and User 4, it did reduce the number of annotated sentences. However, as we have argued previously, annotation time is a more important measure of performance for practical purposes.

Performance of AL and annotation quality. It is evident from the learning curves that annotation quality decreased over time for AL, suggesting that the annotation quality of the more challenging sentences suggested by AL is more vulnerable to the effects of fatigue. Overall, annotation quality for AL of some users (user 1, 3, and 4) was as much as 8% lower than their annotation quality for PL. A more in-depth exploration of this effect is in the Supplementary Material.

DISCUSSION

In this study, we integrated annotation cost estimation models into previously developed AL algorithms and demonstrated the utility of this approach using both simulation and user studies. To the best of our knowledge, Cost-CAUSE is the first AL algorithm to combine uncertainty, representativeness, and cost models to efficiently build NER systems for clinical text. Despite the advantages of the Cost-CAUSE algorithm in this study, these findings suggest there are some aspects of the user annotation process in need of further elucidation.

Differences between users. Our user study showed that the benefit of Cost-CAUSE over PL was different across our 9 users. This is perhaps to be expected, and highlights the danger in drawing conclusions from studies of individuals or pairs of users. Further investigations revealed that changes in annotation quality may explain some of these differences. Also, our annotation cost estimation differed in performance across users (Supplementary Table 3)—especially for User 4, an outlier whose annotation time increased with AL over PL. Clearly it is more difficult to predict the annotation time for some users, and this may cancel out the performance advantage of cost-sensitive methods. Ideally, cost models should work effectively for most users, avoiding the additional annotation associated with individualized cost models. Though out of scope for this study, more features (such as the background of users, time period during an annotation session, reading speed, and other characteristics of users) may be incorporated to develop a future, unified annotation cost model.

Annotation time models. To that end, an interesting research direction may be to develop more sophisticated annotation time models. The relation between annotation time and other annotation time-associated factors may not be adequately represented by a straightforward linear model. For example, interactions between variables such as that between length of the session and length of a sentence (ie, fatigue is more apparent with the more complex AL examples) are currently ignored.

Relation between AL and categories of named entities (NEs). It is possible that AL prefers some categories of entities (see Supplementary Table 7). While the proportions of 3 NEs were similar across users, AL had lower proportions of problem (32.9% vs 38.7%, on average) and higher proportions of test (35.6% vs 29.2% on average) compared to PL for all 9 users. Therefore, incorporating the NE category as a feature might improve the AL method.

Cost-CAUSE in practice. To implement Cost-CAUSE in real-world scenarios, for 1 target data set, we can sample from the data set to train annotators and estimate parameters of the cost model; the remaining samples are used for the main application of AL methods. Because the PL method also needs time to train and test users

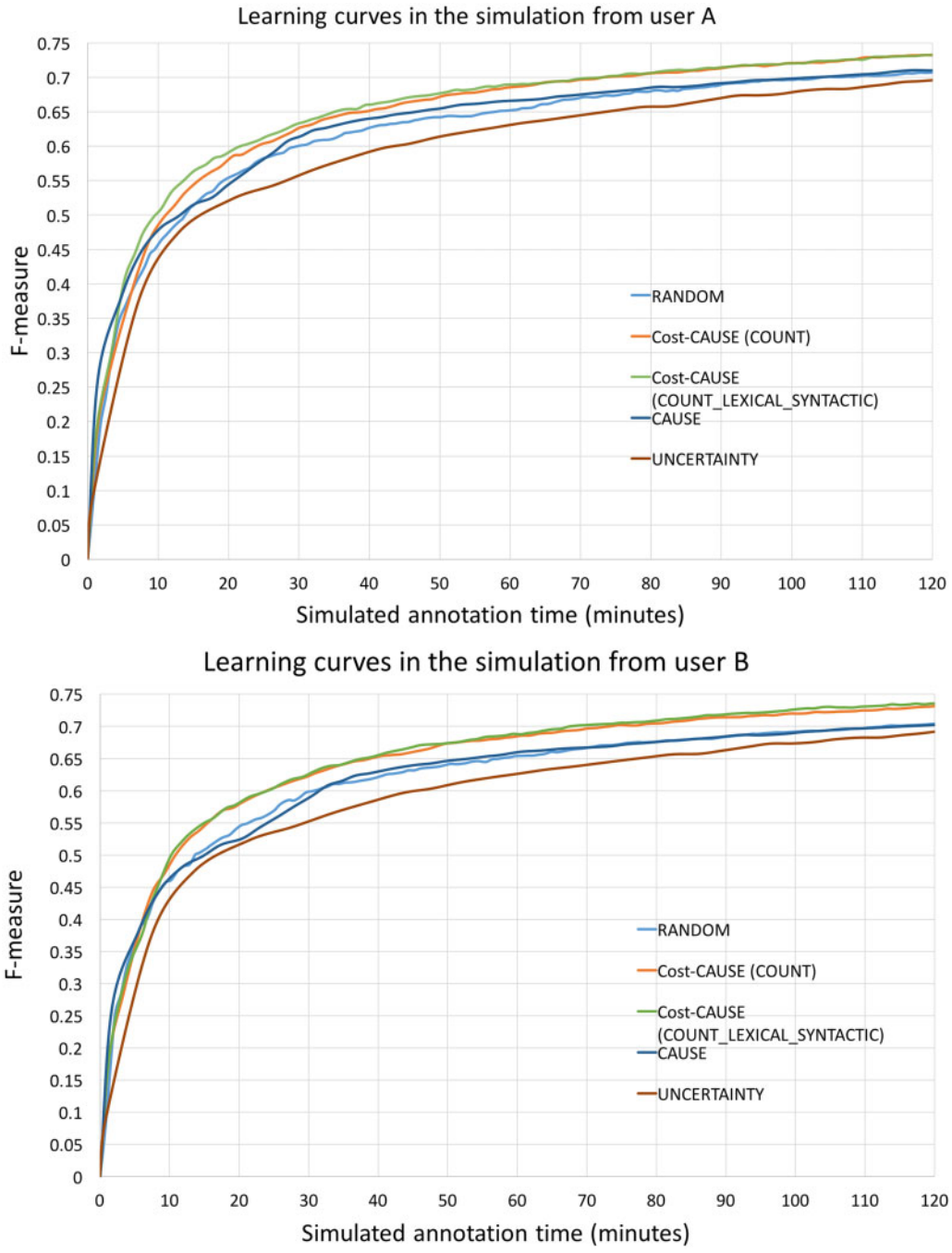

Figure 1. Learning curve for both users in the simulated study, with the simulated annotation time (rather than number of words) on the x-axis. The Cost-CAUSE (COUNT) represents the Cost-CAUSE method that only uses the count features in its cost model, and the Cost-CAUSE (COUNT_LEXICAL_SYNTACTIC) represents the Cost-CAUSE method that uses both lexical and syntactic features in its cost model.

before real annotation, the AL method Cost-CAUSE does not require extra time compared with PL method.

Pros and cons of simulation studies. Aside from using an accurate cost model in AL, the cost model could also be used for simulation studies, which would provide a generalizable evaluation to researchers. Namely, multiple annotation tasks (eg, named entity recognition, word-sense disambiguation, phenotyping) could be evaluated against a simulated annotation time if real annotation time was not explicitly captured. Simulation studies are more economical than user studies when evaluating a large number of methods, and an accurate annotation cost model would enable such simulations to provide estimates of savings in annotator time—the primary outcome of interest from a practical perspective.

However, we should be cautious that simulated results do not overestimate the benefit of AL. There are 2 issues that may cause overestimation of benefit: (1) Simulations are often based on a gold standard (100% quality), whereas a user study relies on annotations generated by users in real time (~80% quality after training). (2) The simulated update process (eg, querying→ annotation→ training→ querying, etc.) is ideal in that a querying can be

Table 3. ALC scores, F-measures at the end of 120-minute annotation, and the statistical test *P* values of AL and PL. Best performance across models for a user is in boldface

Users	ALC scores		F-measures at 120 minutes		*P* values based on Wilcoxon signed-rank test
	PL	AL	PL	AL	
User1	0.633	**0.637**	**0.696**	0.695	$9.7x10^{-2}$
User2	0.574	**0.575**	0.659	**0.671**	$7.2x10^{-3}$
User3	0.608	**0.628**	0.683	**0.690**	$3.0x10^{-5}$
User4	**0.615**	0.609	**0.692**	0.680	$5.6x10^{-3}$
User5	0.619	**0.642**	0.707	**0.717**	$1.8x10^{-5}$
User6	0.580	**0.610**	0.674	**0.691**	$3.9x10^{-4}$
User7	0.521	**0.580**	0.624	**0.671**	$1.8x10^{-5}$
User8	0.599	**0.613**	0.673	**0.691**	$2.7x10^{-5}$
User9	0.606	**0.629**	0.683	**0.693**	$1.8x10^{-5}$
Mean	0.595	**0.632**	0.677	**0.689**	

Abbreviations: AL, active learning; ALC, under the learning curve; PL, passive learning.

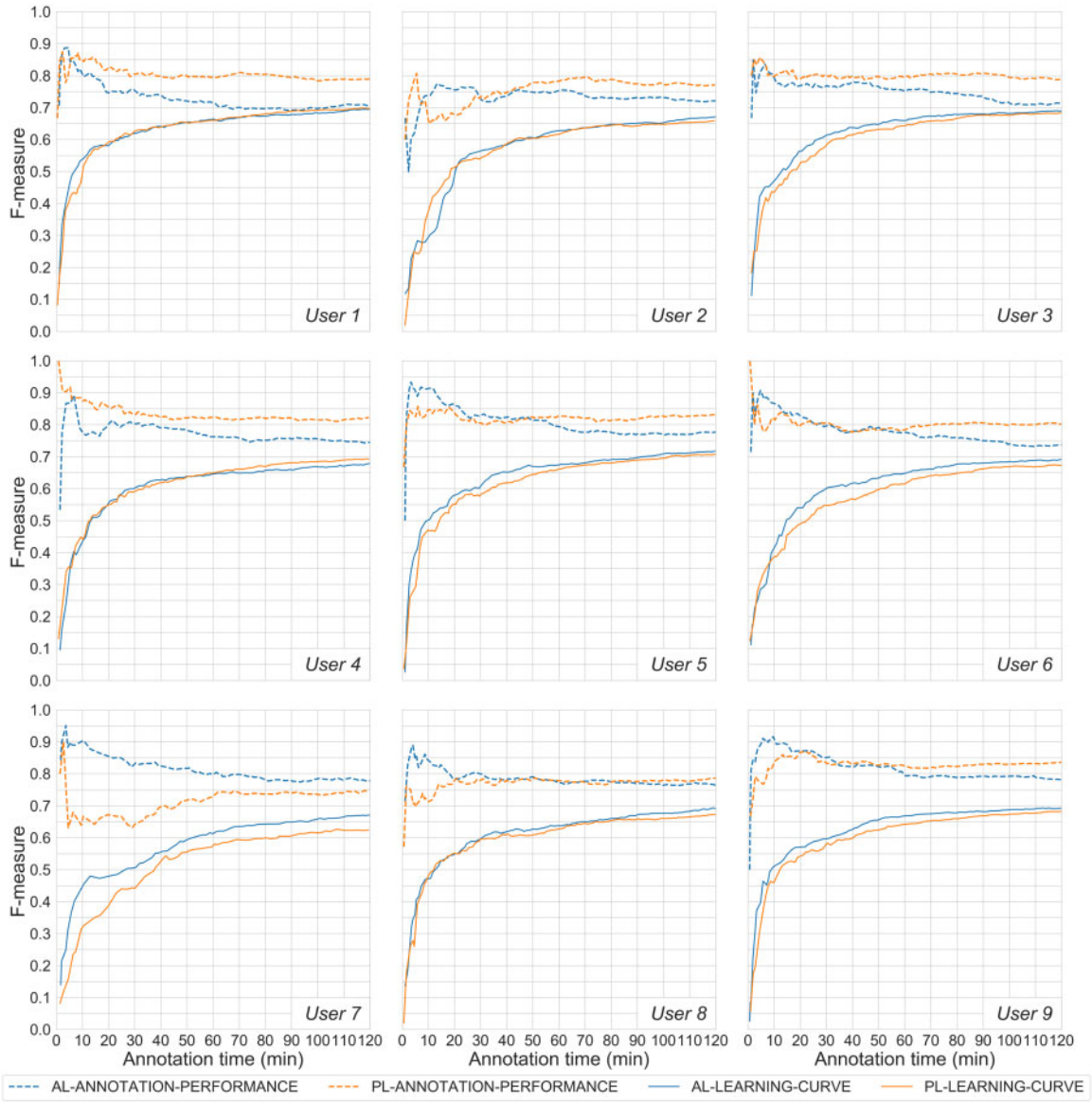

Figure 2. Learning curves and annotation performance for the 9 users. Dashed lines represent annotation quality and solid lines represent the learning curves. The orange and blue represent PL and AL, respectively.

Table 4. Characteristics of annotation processes for 9 users for Random Sampling (PL) and Cost-CAUSE (AL) in each 120-minute annotation

User	Method	Number of sentences	F1	Entities per sentence	Words per sentence	Entity words per sentence	Entity density	Time	Words per minute
User1	PL	920	0.79	1.40	11.37	2.97	0.26	119.97	87.16
	AL	664	0.71	2.06	12.80	4.39	0.34	120.22	70.71
User2	PL	553	0.77	1.44	11.48	2.98	0.26	119.41	53.18
	AL	415	0.72	2.17	12.33	4.50	0.36	120.64	42.43
User3	PL	766	0.79	1.34	10.73	2.74	0.26	119.88	68.58
	AL	525	0.71	1.98	12.16	3.95	0.32	120.19	53.13
User4	PL	842	0.82	1.42	11.64	3.03	0.26	119.56	81.97
	AL	550	0.74	2.38	14.47	4.78	0.33	120.44	66.10
User5	PL	910	0.83	1.43	12.15	3.12	0.26	119.85	92.22
	AL	616	0.78	2.44	14.29	4.88	0.34	120.44	73.07
User6	PL	745	0.80	1.41	11.51	2.99	0.26	120.01	71.44
	AL	570	0.74	2.06	11.20	4.32	0.39	120.22	53.11
User7	PL	435	0.75	1.27	11.11	2.63	0.24	119.46	40.47
	AL	388	0.78	1.86	10.26	3.42	0.33	120.56	33.01
User8	PL	875	0.79	1.38	10.95	2.81	0.26	119.95	79.89
	AL	637	0.77	2.00	11.90	3.83	0.32	120.10	63.10
User9	PL	617	0.84	1.44	11.45	2.93	0.26	120.23	58.77
	AL	445	0.78	2.28	12.52	4.69	0.37	120.04	46.42

Abbreviations: AL, active learning; PL, passive learning.

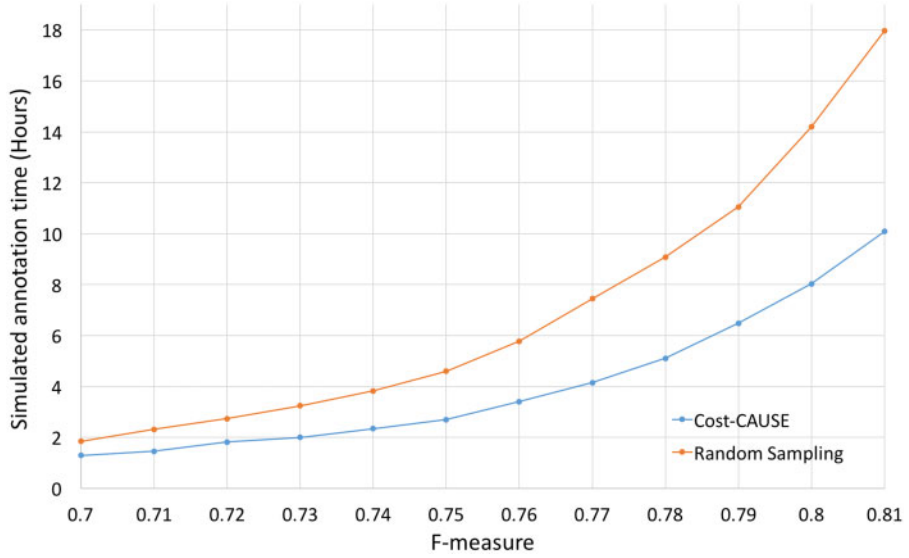

Figure 3. Estimated annotation cost savings by Cost-CAUSE at different F-measures.

performed until the last training is completed, whereas the batch update process in the user study may not be optimal for reasons such as supporting the "no waiting" annotation workflow.

AL for clinical NER in the long term. In this study, we limited the annotation time to 120 minutes, which is not long enough to show the long-term effect of AL methods. To evaluate the long-term performance of AL, we could easily simulate both Cost-CAUSE and random sampling for longer (eg, 20 hours, which is 10 times longer than the user study. Simulated results show that AL would achieve higher percentages of savings when we extend the annotation time, which is very promising (Figure 3). We also plan to extend the user study to evaluate the long-term effect of AL for building clinical NER systems.

User study. Although we developed a system to eliminate bias from factors such as fatigue, memory effect, and so on, limitations to our methods remain. In our study, each sentence could only be selected by either AL or PL, which resulted in different labeled sets for AL vs PL. As previously discussed, annotation quality has an influence on the performance of AL. Participants had diverse backgrounds and natural variability, so it was impossible to ensure higher annotation quality. There are still some factors that may indirectly influence the performance of AL, such as annotation speed. In future work, we will investigate how these direct and indirect factors improve the AL method and define more practical inclusion criterion.

Two dropouts in our 3-phase user study may have some statistical consequences, but they do not affect the outcomes. A single participant drop-out occurred in Phase II due to personal reasons. This could be considered random and thus had no effect on the distribu-

tion of participants' backgrounds. The other dropout was in Phase III due to lower annotation quality, which accorded with the goal of the study—to compare AL and PL in a real environment. Our Phase III AL exclusion criterion was also similar to standard practice for a traditional PL annotation environment, where users undergo training and are only allowed to start annotating real samples if their annotation quality meets the criteria.

CONCLUSION

In this study we presented a cost model to predict annotation time, which was then integrated into the novel AL method, cost-CAUSE. Cost-CAUSE was shown to save annotation time with I2B2 2010 data set in a simulation study. Moreover, we conducted a user study which demonstrated that Cost-CAUSE can save 20.5%–30.2% annotation time in a 2-hour experiment. These results demonstrate the importance of considering, and compensating for, the cost of sentence annotation in clinical AL systems.

FUNDING

This work was supported by National Library of Medicine grant number 2R01LM010681-05.

AUTHOR CONTRIBUTIONS

HX, TC, AF, JD, QM, TL, QC, QW, and YC designed the research. QW, SW, TC, and HX wrote the manuscript. QW, YC, and MS performed the research and analyzed the data under HX, TC, and AF's supervision. All authors revised and approved the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENT

We thank Ky Nguyen, Tolulola Dawodu, and Harish Siddhanamatha for their help in the study.

CONFLICT OF INTEREST STATEMENT

Dr Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

REFERENCES

1. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
2. Liu F, Chen J, Jagannatha A, *et al*. Learning for biomedical information extraction: methodological review of recent advances. *CoRR* 2016;abs/1606.0.http://arxiv.org/abs/1606.07993 (accessed 16 Mar 2017).
3. Kim S, Song Y, Kim K, *et al*. MMR-based active machine learning for bio named entity recognition. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. New York: 2006. 69–72.
4. Chen Y, Lasko TA, Mei Q, *et al*. A study of active learning methods for named entity recognition in clinical text. *J Biomed Inform* 2015; 58: 11–8.
5. Lewis DD, Catlett J. Heterogeneous uncertainty sampling for supervised learning. In: *Machine Learning: Proceedings of the Eleventh International Conference*. San Francisco, CA: 1994. 148–56.
6. Seung HS, Opper M, Sompolinsky H. Query by committee. In: Proceedings of the *Fifth Annual Workshop on Computational Learning* Theory—*COLT* '92. New York, NY: ACM Press; 1992: 287–94.
7. Settles B. Active learning literature survey. Computer Sciences Technical Report 1648. Madison: University of Wisconsin; 2009.
8. Chen Y, Lask TA, Mei Q, *et al*. An active learning-enabled annotation system for clinical named entity recognition. *BMC Med Inform Decis Mak* 2016; 17: 82.
9. Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu: 2008. 1070–9.
10. Kholghi M, Sitbon L, Zuccon G, *et al*. Active learning: a step towards automating medical concept extraction. *J Am Med Inform Assoc* 2015;23:289–96.
11. Settles B, Craven M, Friedland L. Active learning with real annotation costs. In: *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*. Vancouver, CA: 2008. 1–10.
12. Kholghi M, Sitbon L, Zuccon G, *et al*. Active learning reduces annotation time for clinical concept extraction. *Int J Med Inform* 2017; 106: 25–31.
13. Haertel RA, Seppi KD, Ringger EK, Carroll JL. Return on investment for active learning. In: *Proceedings of the Neural Information Processing Systems Workshop on Cost Sensitive Learning 2008*. Vancouver, CA: 2008. Vol. 72.
14. Tomanek K, Hahn U. A comparison of models for cost-sensitive active learning. In: *Proceedings of the 23rd International Conference on Computational Linguistics*; Beijing: 2010. 1247–55.
15. Ringger E, Carmen M, Haertel R, *et al*. Assessing the costs of machine-assisted corpus annotation through a user study. In: *LREC 2008*. Marrakech: 2008. 3318–24.
16. Arora S, Nyberg E, Rosé CP. Estimating annotation cost for active learning in a multi-annotator environment. In: *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*. Boulder, Colorado: 2009. 18–26.
17. Tomanek K, Hahn U, Lohmann S, *et al*. A cognitive cost model of annotations based on eye-tracking data. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: 2010. 1158–67.
18. Haertel R, Ringger E, Seppi K, *et al*. An analytic and empirical evaluation of return-on-investment-based active learning. In: *Proceedings of The 9th Linguistic Annotation Workshop*. Stroudsburg, PA: 2015. 11–20.
19. Uzuner Ö, South BR, Shen S, *et al*. i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
20. Wu S, Bachrach A, Cardenas C, *et al*. Complexity metrics in an incremental right-corner parser. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: 2010. 1189–98.
21. Kuperberg GR, Jaeger TF. What do we mean by prediction in language comprehension? *Lang Cogn Neurosci* 2016; 31 (1): 32–59.