Active Learning for Clinical Named Entity Recognition

Hua Xu, PhD

School of Biomedical Informatics

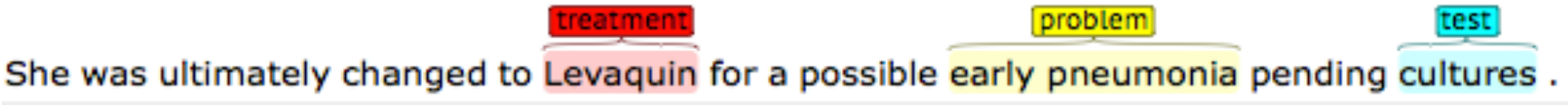
The University of Texas Health Science Center at Houston

Disclosure

- Founder of Melax Technologies Inc
- Consultant of Hebta LLC, More Health, DCHealth Technologies Inc.

Named entity recognition (NER)

- NER is a fundamental task in clinical NLP



- Rule-based NER
  - Dictionary lookup
  - Regular-expression rules
- Machine learning (ML)-based NER
  - Growth of large annotated datasets
  - Shown better performance in multiple NLP challenges

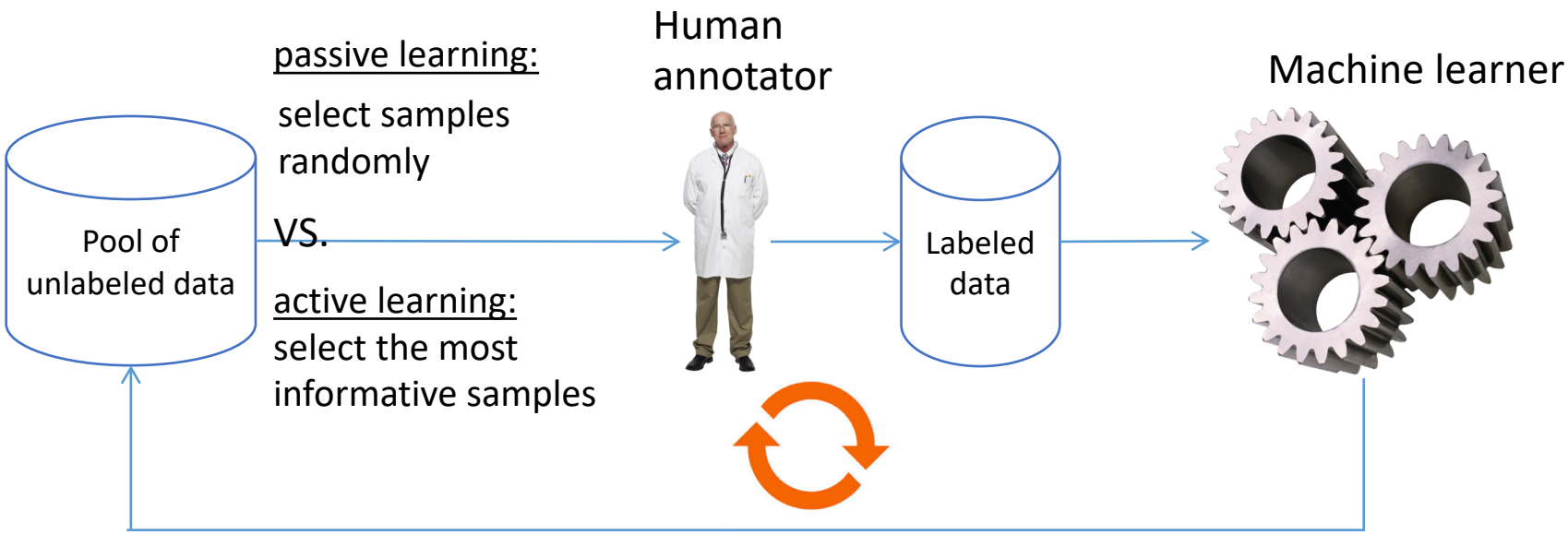
Challenges of ML-based NER

- Large annotated corpora are expensive to build
  - Need domain experts
  - Time consuming
- ML models are not generalizable
  - From one type of clinical notes to another
  - From one institution to another

Minimize annotation cost while optimizing ML-based models

Active learning

- Goal: minimize annotation cost while maximizing the quality of ML-based model



Active learning for clinical NER

- Few studies focused on clinical NER – a sequence labeling problem
  - More difficult than binary classification
  - More complex in measuring the value of sample

- Most previous studies used simulation
  - Not consider real annotation cost
  - Not apply AL to annotation in practice

Active learning methods for clinical NER using simulation

A clinical NER task

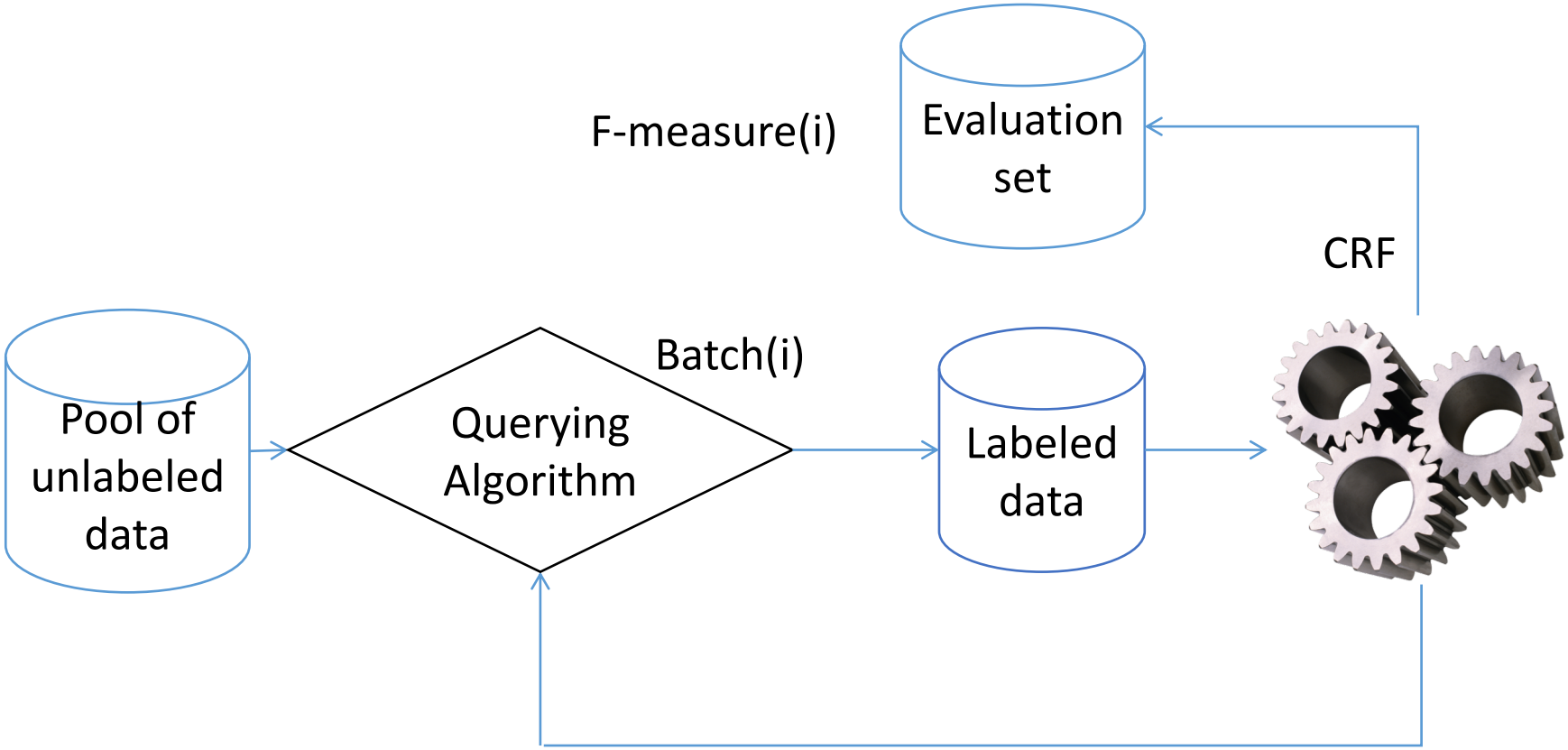
- To automatically identify three categories of clinical concepts/entities in clinical notes
  - Problems
  - Treatments
  - Lab tests
- ML-based NER modeling
  - Sequential labeling algorithm: conditional random field (CRF)
  - Features: word, syntactic, and semantic levels
  - Assign a label (BIO) for each word in the sentence

She	was	ultimately	changed	to	Levaquin	for	a	possible	early	pneumonia	pending	cultures	.
O	O	O	O	O	B-treatment	O	O	O	B-problem	I-problem	O	B-test	O

Dataset

- 349 annotated clinical notes from 2010 i2b2/VA NLP challenge
- 20,423 unique sentences
- 5-fold splits
  - ~16,338 sentences in the pool
  - ~4,085 sentences in the evaluation set

Active learning simulation framework



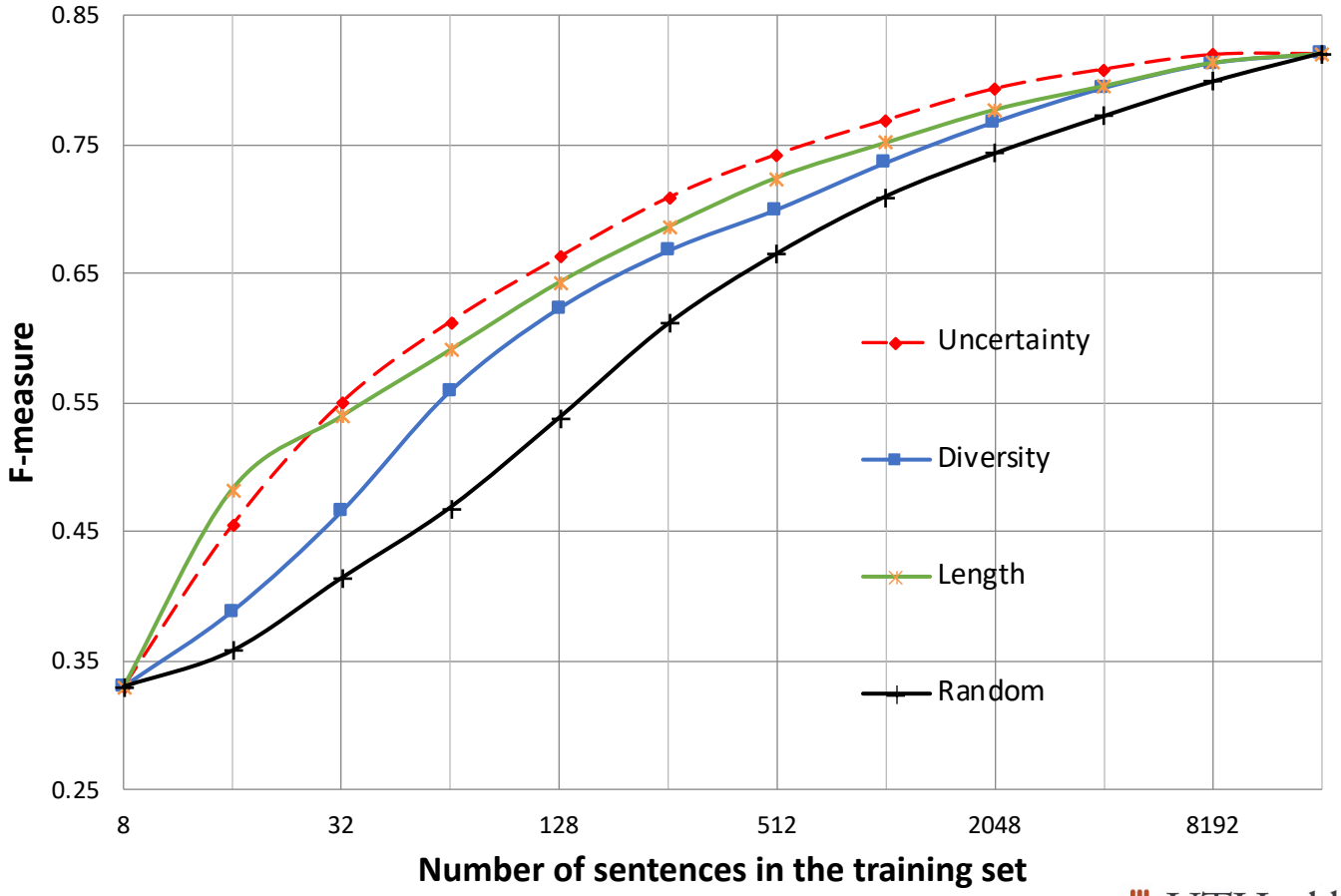
Querying algorithms

- Uncertainty sampling
  - Query the sentences with MOST uncertainty
  - 6 uncertainty measurements
- Diversity sampling
  - Query the sentences LEAST similar to those already annotated
  - 4 similarity measurements
- Baselines
  - Length – words
  - Length – concepts
- Random sampling

Evaluation

- Annotation Cost
  - Number of sentences in the training set
- Learning curves and ALC (area under the learning curve) score
  - ALC: F-measures vs. sentences
- 5-fold cross validation
  - Final learning curve = the average of 5 learning curves

Learning curves (same cost per sentence)



Result highlights

	ALC (F-measure vs. sentences
Uncertainty sampling	0.83
Diversity sampling	0.79
Length	0.82
Random sampling	0.74

To achieve a model with 0.80 in F-measure

Annotation cost	Random sampling	Uncertainty sampling	Reduction percentage over Random
Sentences (Traditional)	8,702	2,971	66%

A problem of uncertainty sampling

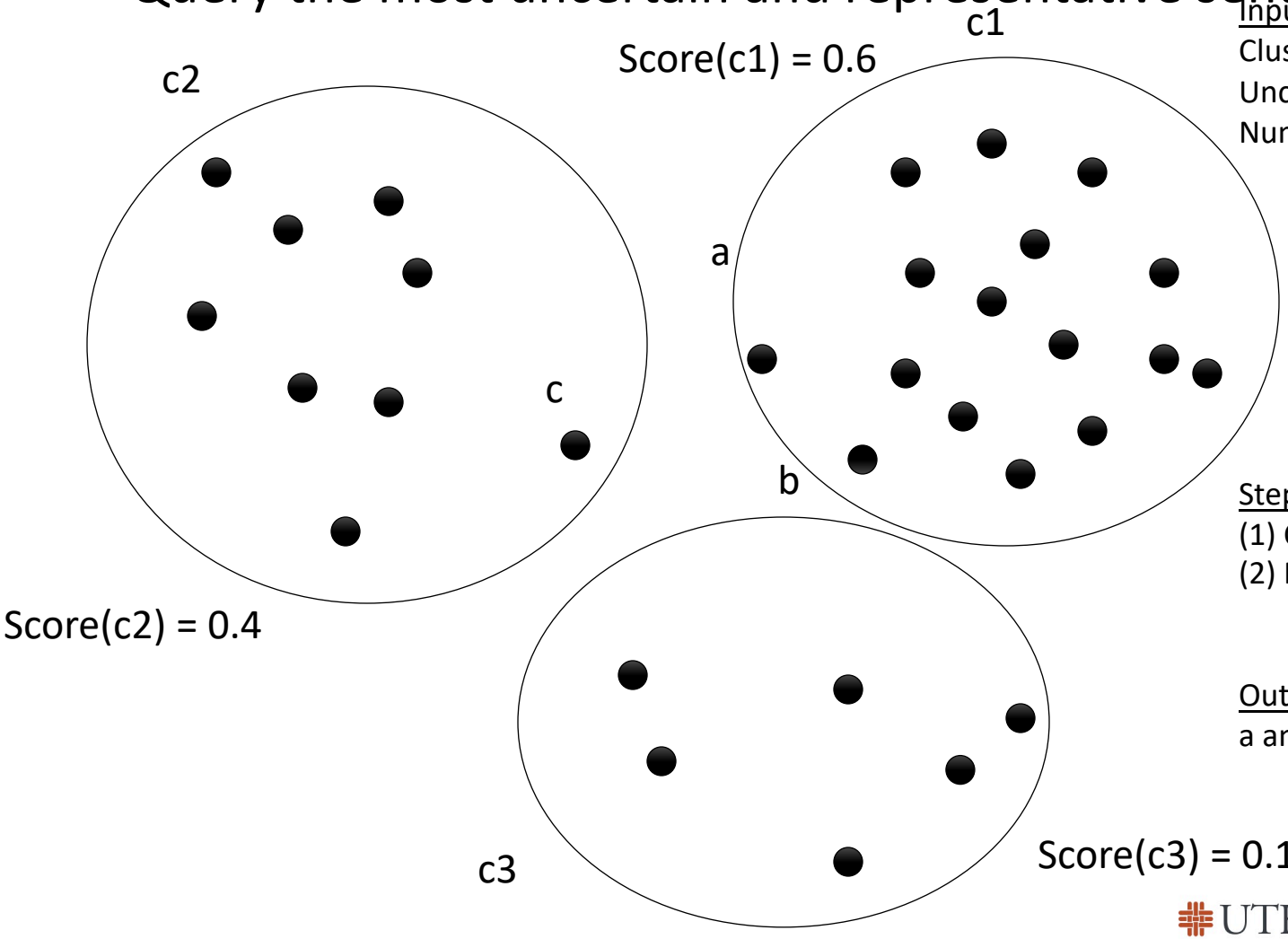
- Similar sentences occurred back-to-back in a trial run using uncertainty sampling

Coronary Artery Disease, Hypertension, Hyperlipidemia, Diabetes Mellitus, Hypothyroid, h/o Bilateral DVT's (on chronic coumadin therapy), Pleural disorder? Sarcoidosis, Gastritis, B12 deficiency, Chronic renal insufficiency, s/p Appendectomy, s/p Lap cholectomy, s/p Total abdominal hysterectomy

PMH: Hypertension, Hyperlipidemia, Diabetes Mellitus, Hypothyroid, h/o Bilateral DVT's, Pleural disorder? Sarcoidosis, Gastritis, B12 deficiency, Chronic renal insufficiency, s/p Appendectomy, s/p Lap cholectomy, s/p Total abdominal hysterectomy

Clustering and uncertainty sampling engine (CAUSE)

- Query the most uncertain and representative sentences

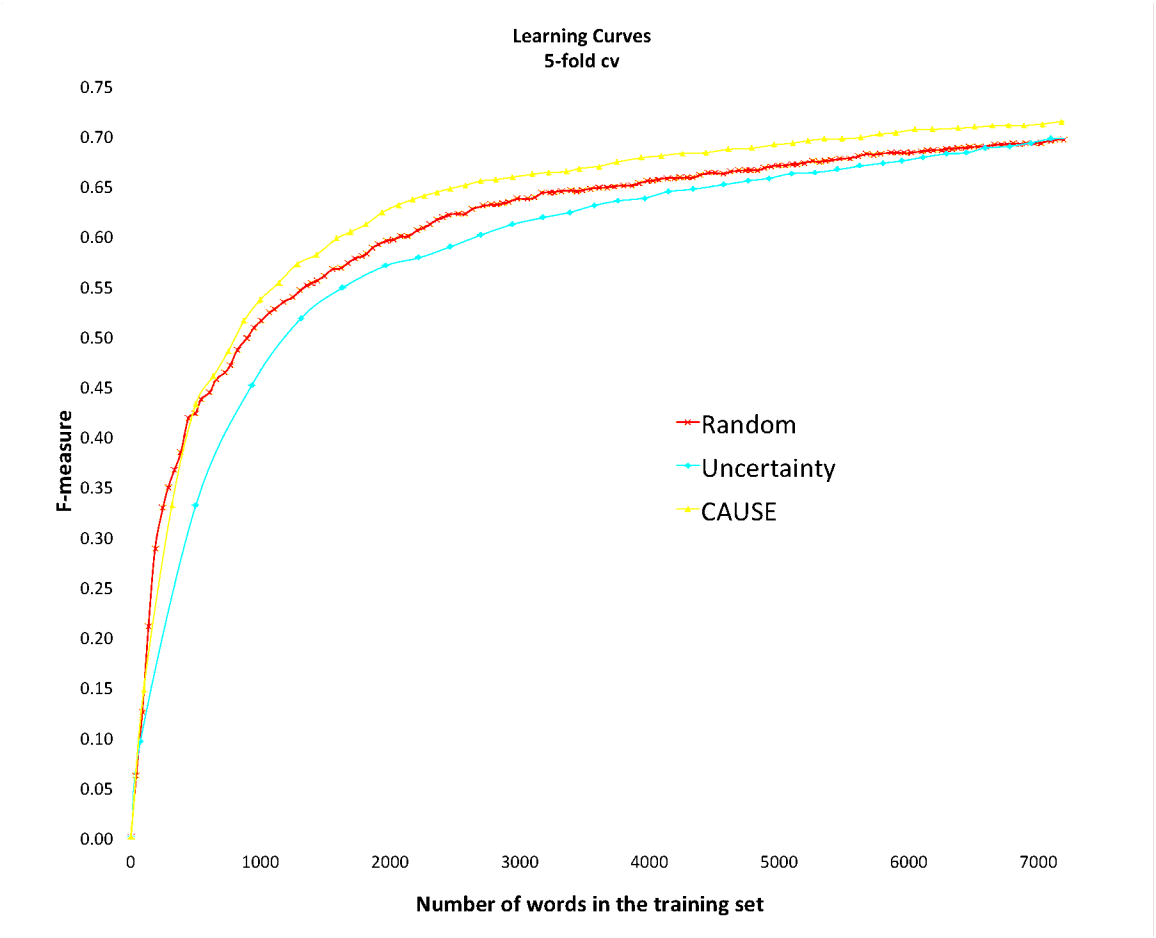


Inputs:
Clusters
Uncertainty
Number of queries = 2

Steps:
(1) Cluster scoring
(2) Representative sampling

Outputs:
a and c

Simulation Experiment results

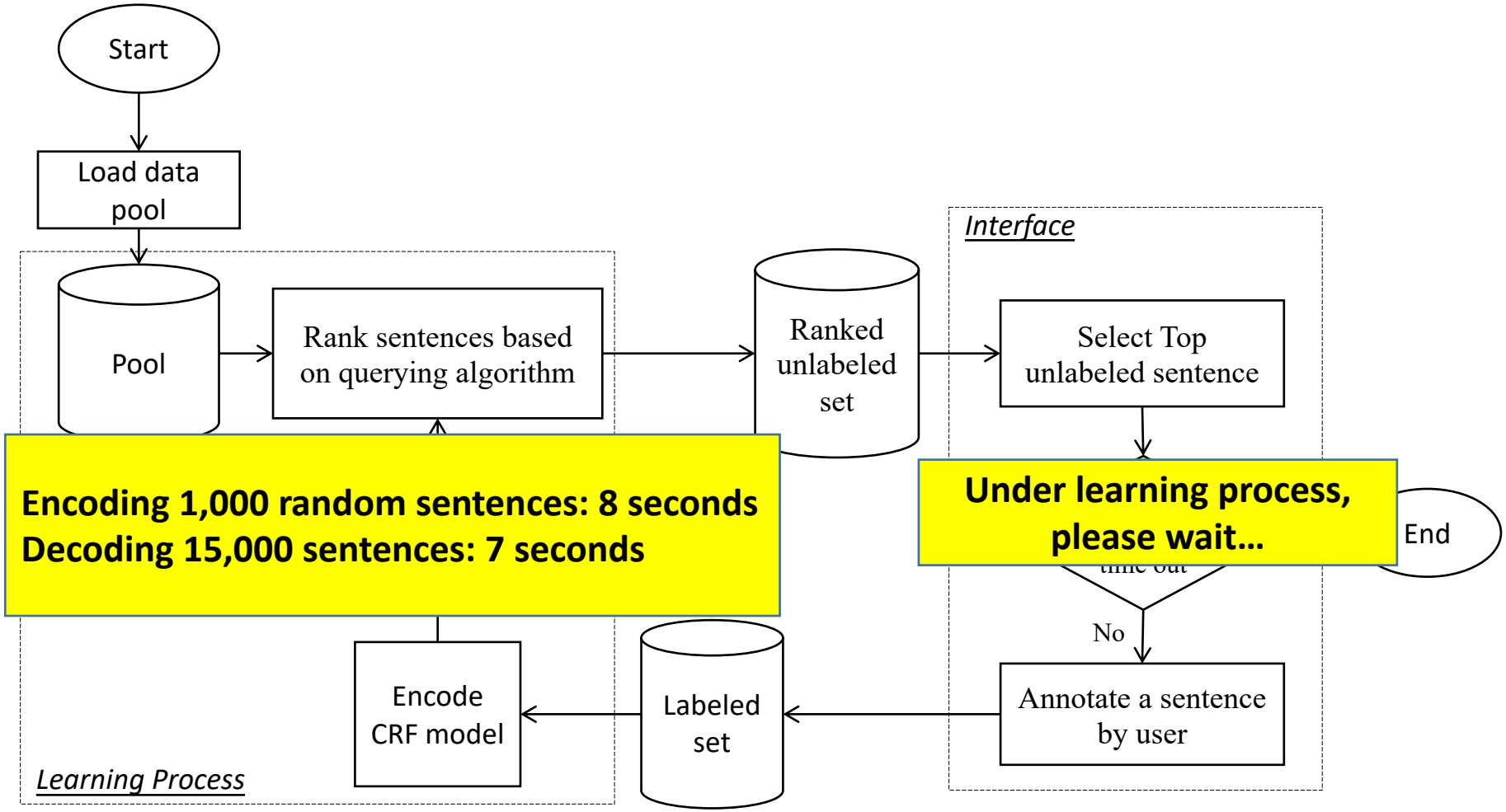


Build and evaluate an active learning-enabled annotation system for clinical NER in practice

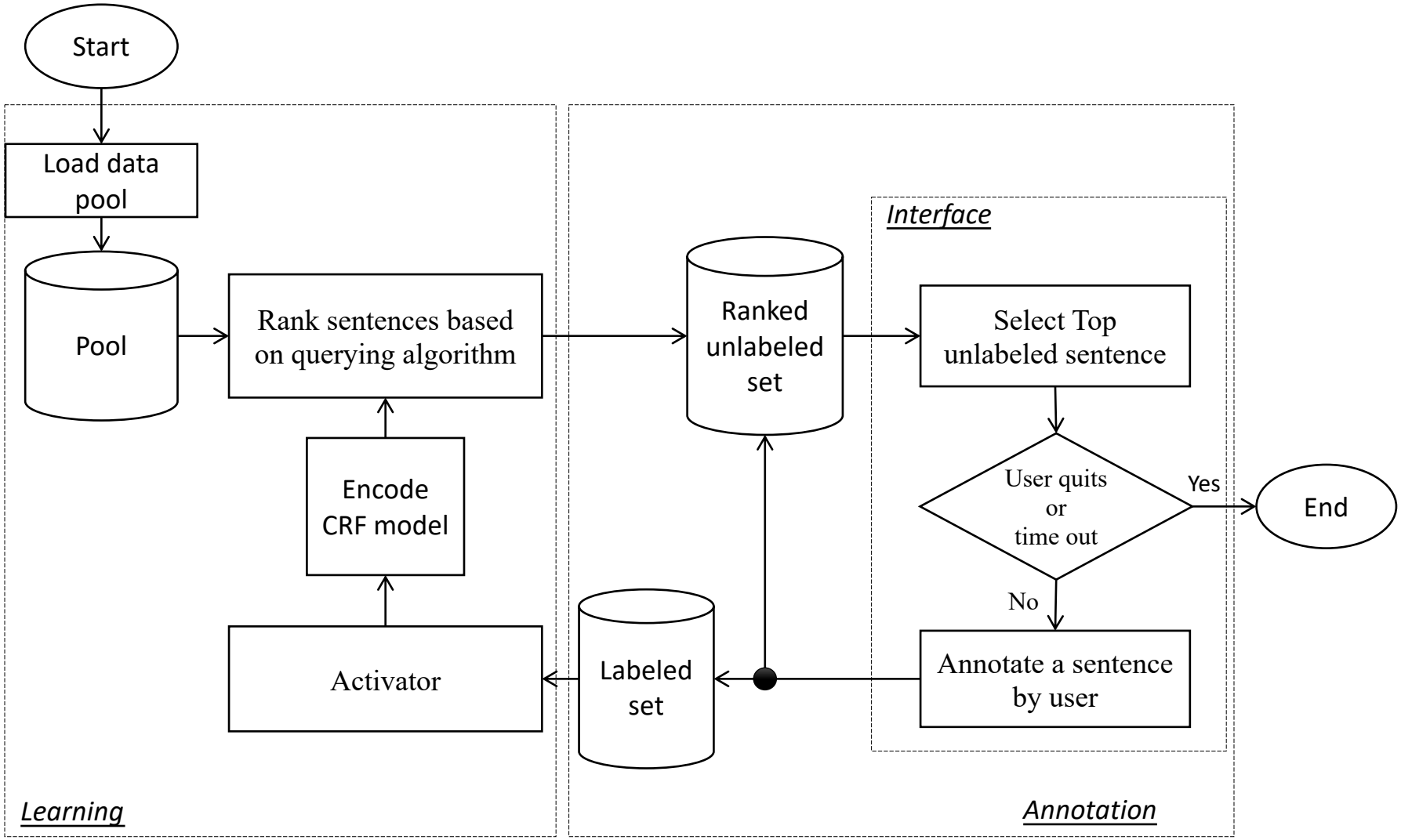
Active LEARNER

- Active Learning-Enabled AnnotatoR for NER
- Three components
  - Annotation interface
  - Machine learning engine
  - Querying engine
- Implementation details
  - Java Eclipse based application
  - BRAT annotation system
  - CRF ++ for NER

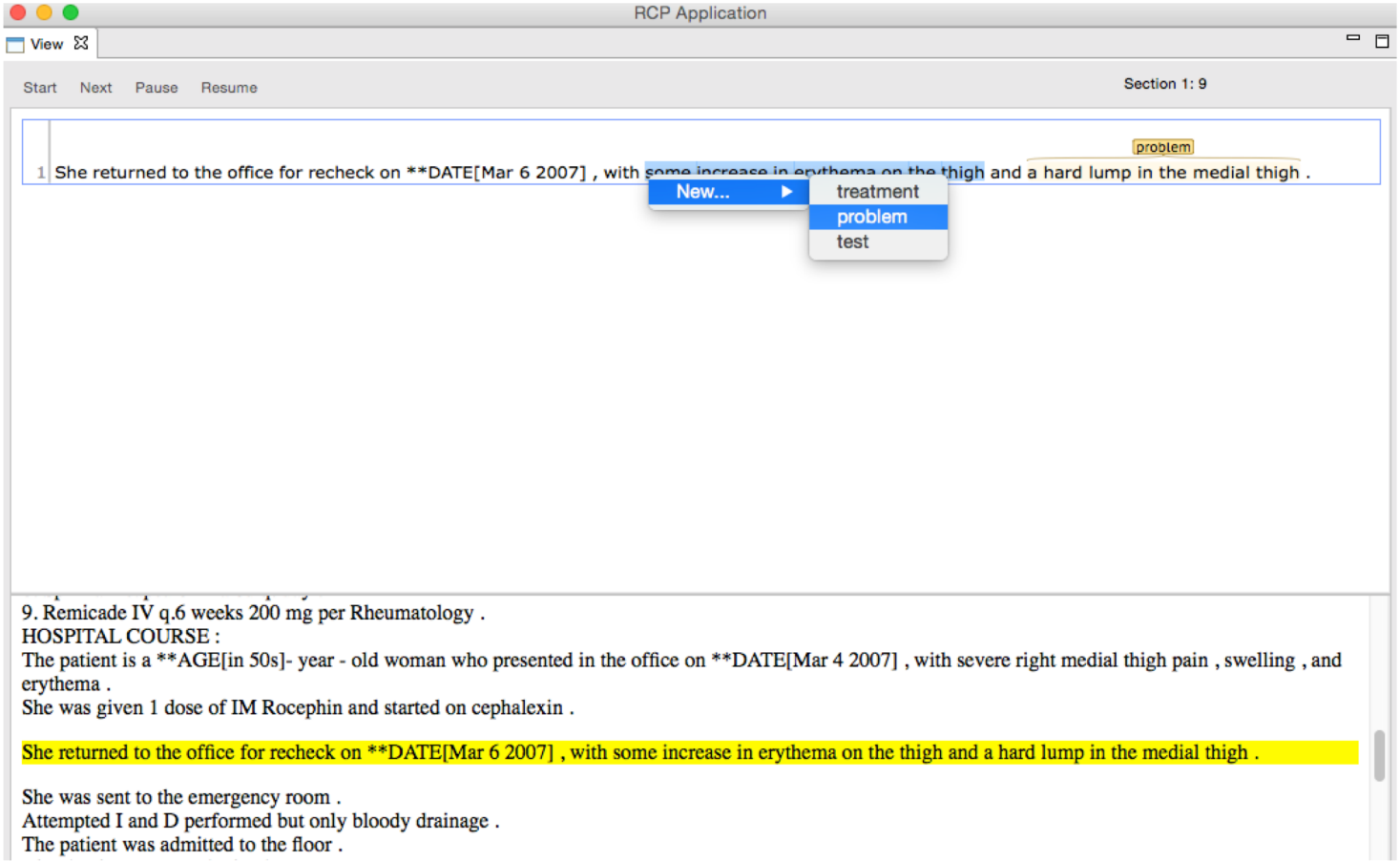
Initial workflow



Final workflow



Hybrid active learning system



Annotation interface of active learning system.

User study design

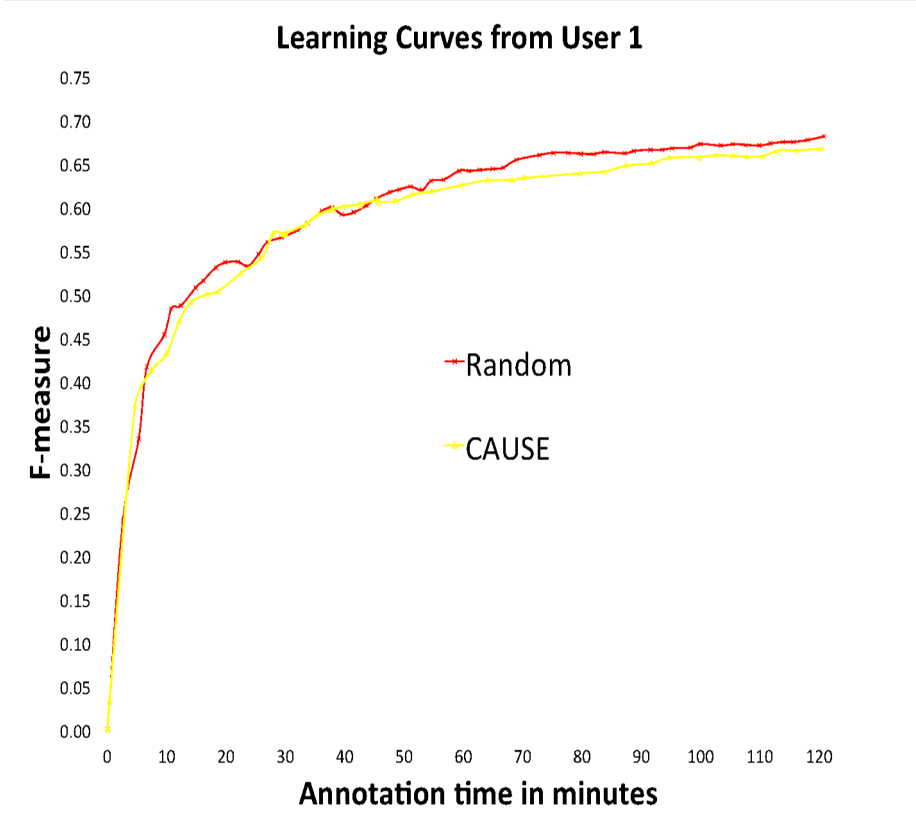
- Two nurses were recruited

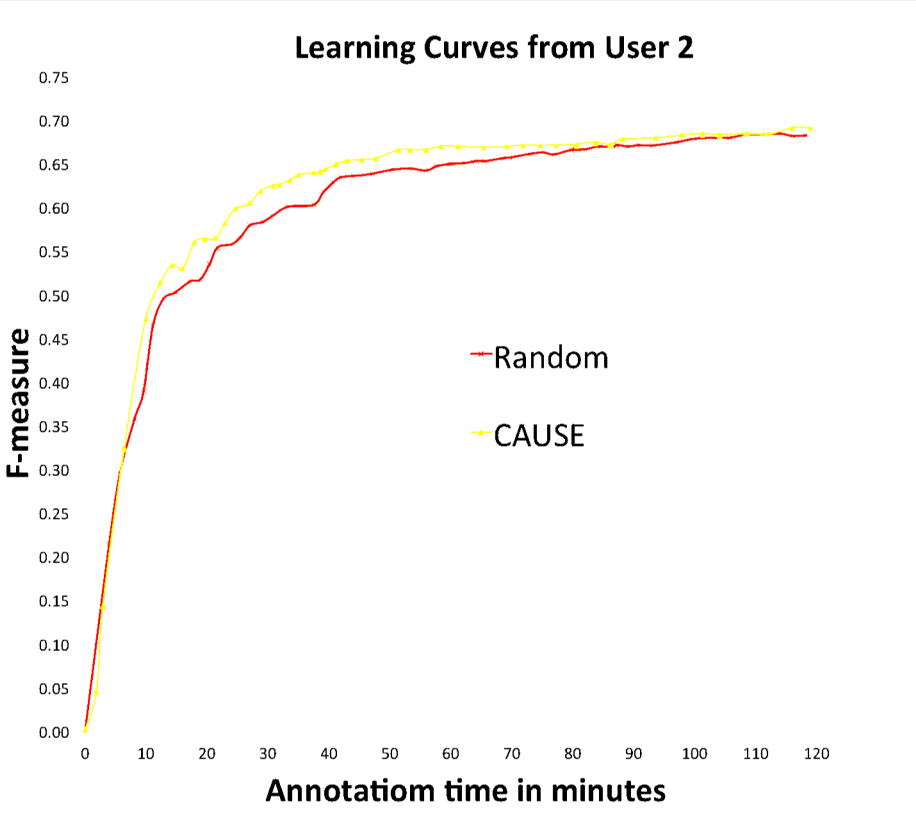
Time	Event	Task	Duration
Week 0	**Guided Training**	1. Annotation guidelines review	30 minutes
		2. Sentence-by-sentence annotation review	45 minutes
	Practice	1. Three 15-minute sessions of annotation practice	45 minutes
		2. Four 30-minute sessions of annotation using *Random*	3 hours
Week 1	**Annotation warm up training**	1. Sentence-by-sentence annotation review	15 - 30 minutes
		2. Two 15-minute sessions of annotation practice	30 minutes
	Main study (*Random*)	Four 30-minute sessions of annotation	3 hours
Week 2	**Annotation warm up training**	1. Sentence-by-sentence annotation review	15 - 30 minutes
		2. Two 15-minute sessions of annotation practice	30 minutes
	Main study (*CAUSE*)	Four 30-minute sessions of annotation	3 hours

Evaluation of user study

- Dataset:
  - 16,338 sentences in the pool
  - 4,085 sentences in the evaluation set
- Learning curves
  - F-measures vs. real annotation time
- Users' annotation performance
  - Annotation speed (annotated entities per minute)
  - Annotation quality (F-measure against gold standard)

User study results: Learning curves from both users





User study result highlights

Users	Methods	ALC scores	F-measure of models at 120 minutes
User 1	*Random*	0.81	0.68
	CAUSE	0.78	0.67
User 2	*Random*	0.82	0.68
	CAUSE	0.83	0.69

User	Methods	Annotation speed (entities per minute)	Annotation quality (F-measure)
User 1	*Random*	7.88	0.82
	CAUSE	7.72	0.83
User 2	*Random*	7.35	0.81
	CAUSE	7.90	0.82

What we learned from the user study?

- AL samples are more difficult than random samples
- Users' annotation behaviors are different
- User response (i.e., speed, quality) is changing during the annotation due to different factors such as fatigue
- Order of AL vs. PL – does wash out period work?
- ….

Annotation time for each sentence is different – we have to consider the annotation cost (time) when selecting samples

- Settles et al. reported an empirical study of AL with real annotation costs
  - When annotation cost per sample varies, AL without cost model performed NO better than random sampling
  - AL with cost variable appropriately considered could be improved in some cases
- Haertel et al. presented a cost-conscious AL based on return on investment (ROI)
  - Applied ROI active learning in part-of-speech tagging
  - Saved as high as 73% in hourly cost

Cost-aware AL for clinical NER

Utility per Cost (UPC) model

A linear regression model was used to estimate annotation time based on the basic information, semantic complexity and syntactic complexity of the sentence.

$$\text{UPC(s)} = \frac{utility(s)}{cost(s)}$$

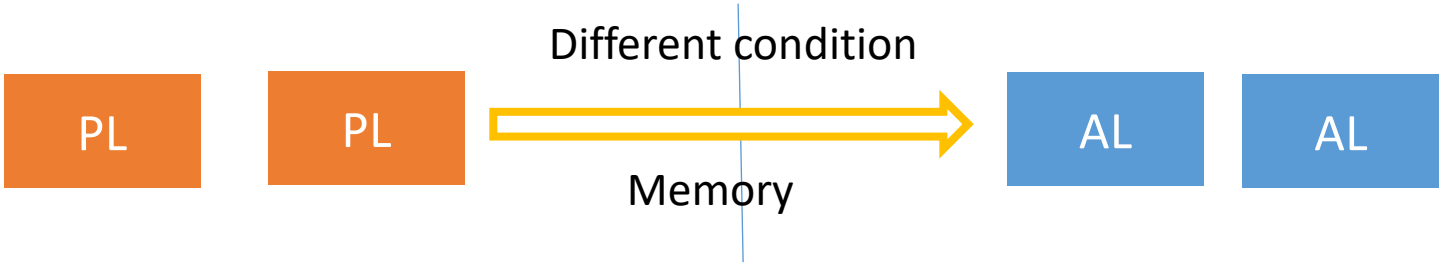
$$cost(s) = c_0 + \sum_i c_i f_i(s)$$

Categories	Features
Basic	Number of words (NOW)
	Number of entities (NOE)
	Number of entity words (NOEW)
Syntactic	Entropy of POS tag (EOP)
Semantic	TFIDF

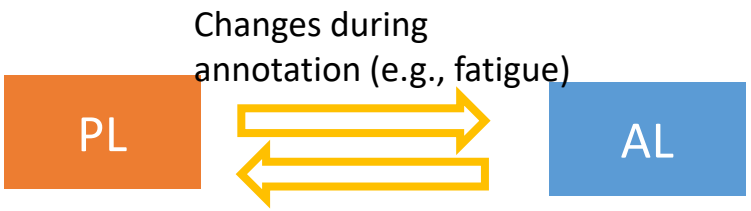
Sentence	***MRI*** by report showed ***bilateral rotator cuff repairs*** and he was admitted for ***repair of the left rotator cuff*** .				
Feature	NOW	NOE	NOEW	TFIDF	EOP
Value	20	3	11	35.36	2.28

How to design the experiment to compare AL and PL fairly?

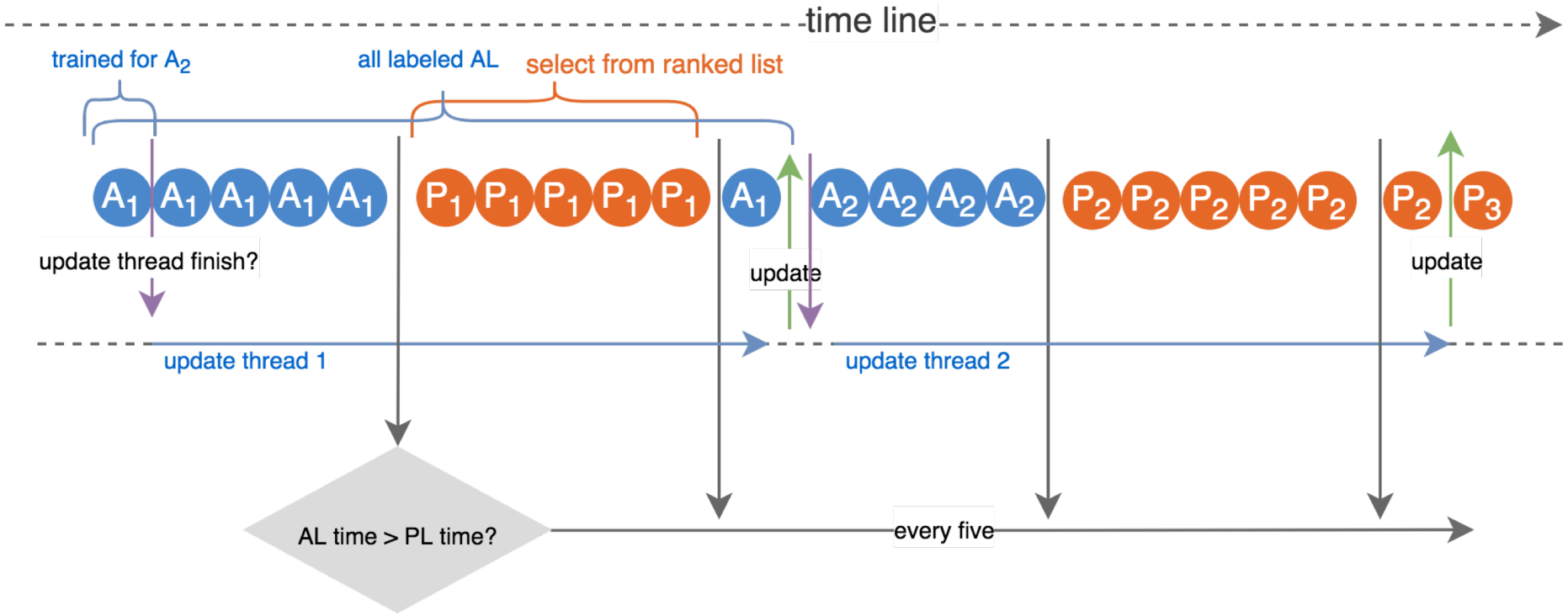
1. Annotate with a wash out period



2. Annotate on the same day.



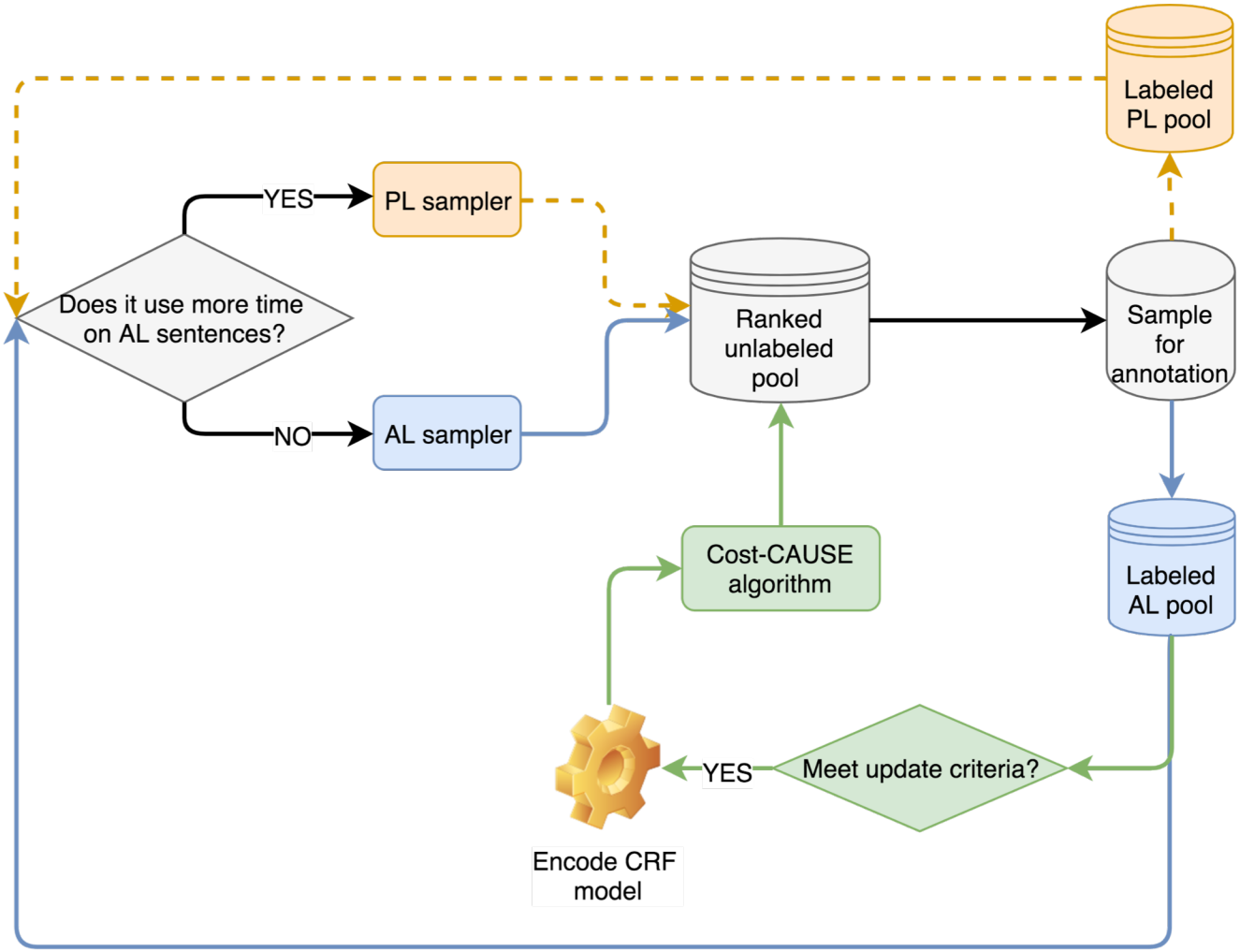
A hybrid design



A hybrid system for comparison AL and PL.

1. The sentences from AL and PL are provided to user alternatively. The user doesn't know how sentences are sampled.
2. It doesn't need to stop when the ML model is trained.
3. It keeps user spends same time on AL and PL.

Workflow of hybrid system

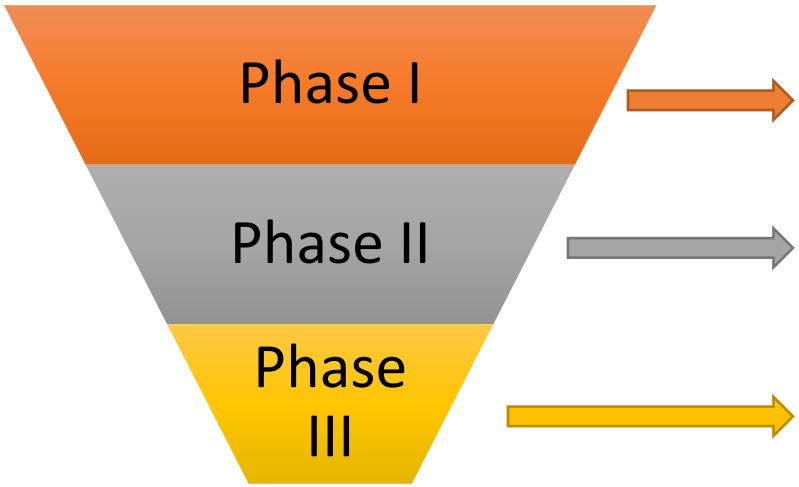


A hybrid system can select sentences with AL and PL method alternatively and ensure user spends same time on AL and PL.

A larger user study

Inclusion condition:

1. Individual who have medical training and have worked with clinical notes written in English
2. Medical students (> second year)
3. Nursing students (> first year)



- ✓ **Interview** - One hour, 20 users.
- ✓ Learn what's annotation and basic test.

- ✓ **Training** - Four hours, 12 users, two days.
- ✓ Learn guideline, practice on the system and review their annotation. At the end, take one hour test .

- ✓ **Main study** - Six hours, 10 users, two days. (1 user was removed from further analysis due to lower annotation quality)
- ✓ Six sessions of annotation (40 min/session). **User annotates by AL and PL for 120 minutes, respectively.**

Data in the study

- I2b2/VA 2010 dataset
  - Training: 349 clinical documents with 20,423 unique sentences.
  - Test: 477 clinical documents with 29,789 unique sentences.
- Three types of medical entities: problem, treatment, and test
- User study
  - Phase I : training corpus.
  - Phase II : training corpus
  - Phase III : test corpus

Result - Analysis of annotation for AL and PL

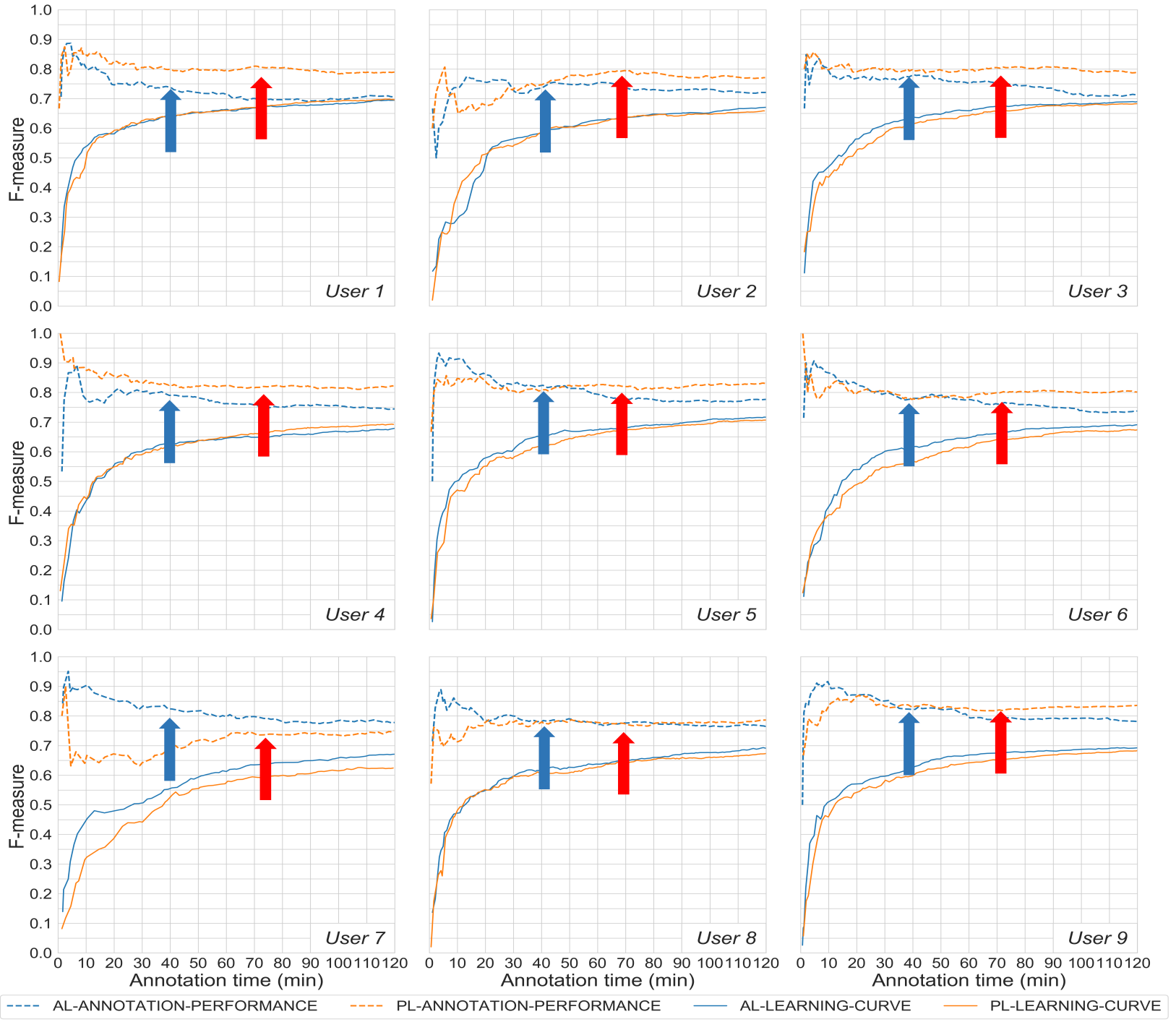
	AL	PL
Number of sentences	534±**100**	740±**170**
Words per sentence	12.44±1.34	11.38±0.41
Entities per sentence	2.14±0.20	1.39±0.06
Entity density	0.34±0.02	0.26±0.01
Speed (words/min)	55.7±**13.6**	70.4±**17.0**
F1 for annotation	74.8±0.03	79.8±0.03

The annotation of nine users in the user study were analyzed. The sentences selected by AL contained more entities than those by PL, which took longer time for users to annotate.

Result - ALC score for AL and PL

Users	ALC scores		F-measures at 120 minutes		P-values based on Wilcoxon signed-rank test
	AL	PL	AL	PL	
User1	0.6370	0.6326	0.6947	0.6963	9.7×10^{-2}
User2	0.5749	0.5740	0.6710	0.6585	7.2×10^{-3}
User3	0.6276	0.6083	0.6904	0.6827	3.0×10^{-5}
User4	0.6089	0.6151	0.6799	0.6923	5.6×10^{-3}
User5	0.6422	0.6190	0.7166	0.7068	1.8×10^{-5}
User6	0.6103	0.5799	0.6913	0.6740	3.9×10^{-4}
User7	0.5799	0.5209	0.6712	0.6242	1.8×10^{-5}
User8	0.6131	0.5992	0.6911	0.6731	2.7×10^{-5}
User9	0.6285	0.6055	0.6925	0.6827	1.8×10^{-5}

The ALC for AL is higher than that for PL in the eight of nine users.


F-measure
Annotation time (min)
User 1
User 2
User 3
User 4
User 5
User 6
User 7
User 8
User 9
AL-ANNOTATION-PERFORMANCE
PL-ANNOTATION-PERFORMANCE
AL-LEARNING-CURVE
PL-LEARNING-CURVE

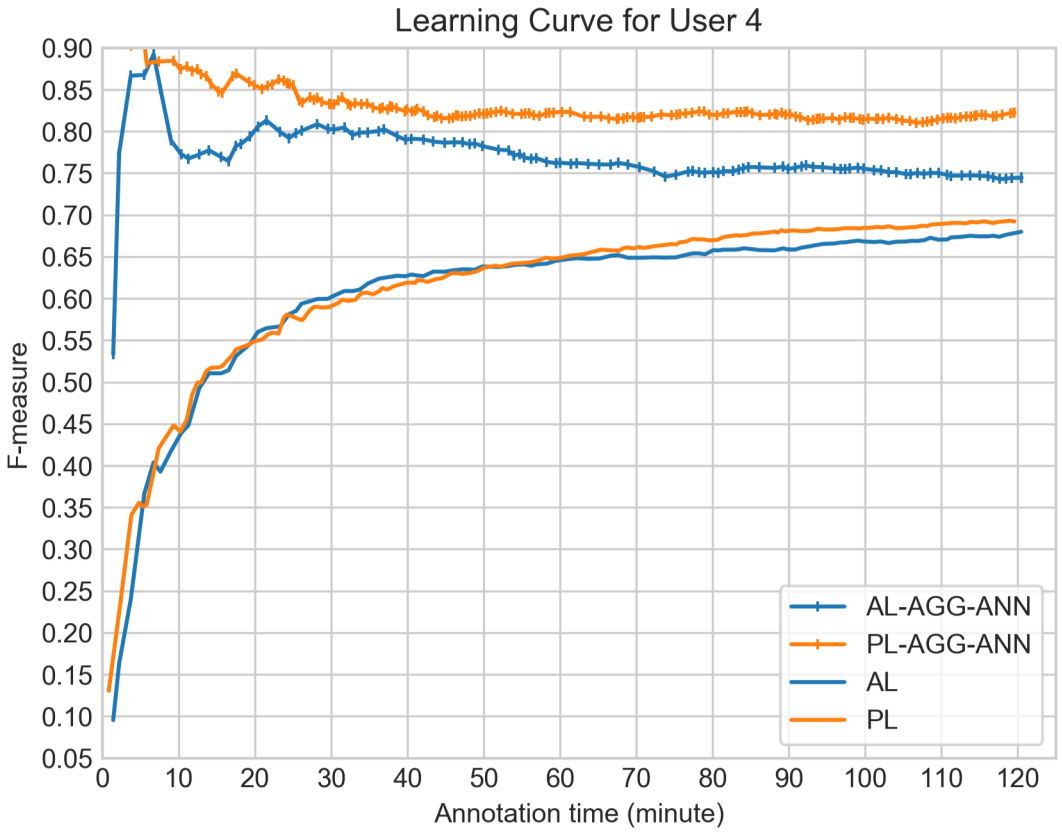
Result - Reduction of annotation cost

	Time (min)		Reduction %	#Sentences		Reduction %	#Words		Reduction %
Method	AL	PL		AL	PL		AL	PL	
User4	108.5	77.3	-40.3	490	535	8.4	7147	6065	-17.8
User1	71.2	70.2	-1.5	375	515	27.2	4734	5964	20.6
User3	66.3	83.4	20.5	270	495	45.5	3434	5501	37.6
User8	90.3	117.1	22.9	465	850	45.3	5530	9282	40.4
User6	75.6	101.7	25.6	345	605	43	3907	7029	**44.4**
User5	48.2	65.9	26.8	220	435	**49.4**	3318	5500	39.7
User9	64.2	91.9	**30.2**	255	465	45.2	3247	5309	38.8
User2	120.6	-	-	415	-	-	5119	-	-
User7	117.3	-	-	375	-	-	3884	-	-

*A F-measure of 0.67 was chosen to compare reduction of annotation cost measured by time, words and sentences for AL and PL.

Although saved annotation effort measured by reduction of number of sentences and words are large, saved effort measured by annotation time is not as much as as them in practice.

Discussion - Relation between AL performance and annotation quality



	PCC	P value
User1	0.7508	2.37e-05
User2	-0.6902	1.90e-04
User3	0.8420	2.50e-07
User4	0.9550	4.27e-13
User5	0.9104	6.79e-10
User6	0.1646	4.42e-01
User7	0.5124	1.05e-02
User8	0.4651	2.20e-02
User9	0.6418	7.23e-04

$\textit{tation\_quality\_PL}_i$

$$\Delta \text{model}\_performance_i = model\_performance\_AL_i\text{-}model\_performance\_PL_i$$

To explore whether difference between annotation quality for AL and PL had an effect on relative performance of generated ML model by AL and PL, we calculated the Pearson correlation coefficient between them.

Discussion

- Limitation
  - Annotation quality of some users for AL is low.
  - The total annotation time for AL and PL (two hours) are not long enough to show the long-term performance of them.
  - The cost model is specific to individual and performance of cost model has a difference across users (0.56 – 0.87, R^2).
- Future
  - Improve the training method and extend the training time.
  - Conduct a longer time user study.
  - Develop a cost model that can be used for all users and includes more variables like characteristics of users.

Acknowledgement

- UTHealth
  - Qiang Wei
  - Trevor Cohen
  - Amy Franklin
  - Anupama Gururaj

Vanderbilt

Yukun Chen, PhD

Joshua C. Denny, MD, MS

Thomas A. Lasko, MD, PhD

Qingxia Chen, PhD

- University of Michigan
  - Qiaozhu Mei, PhD

Grant:

NLM R01 LM010681

Thank you!

Hua.xu@uth.tmc.edu