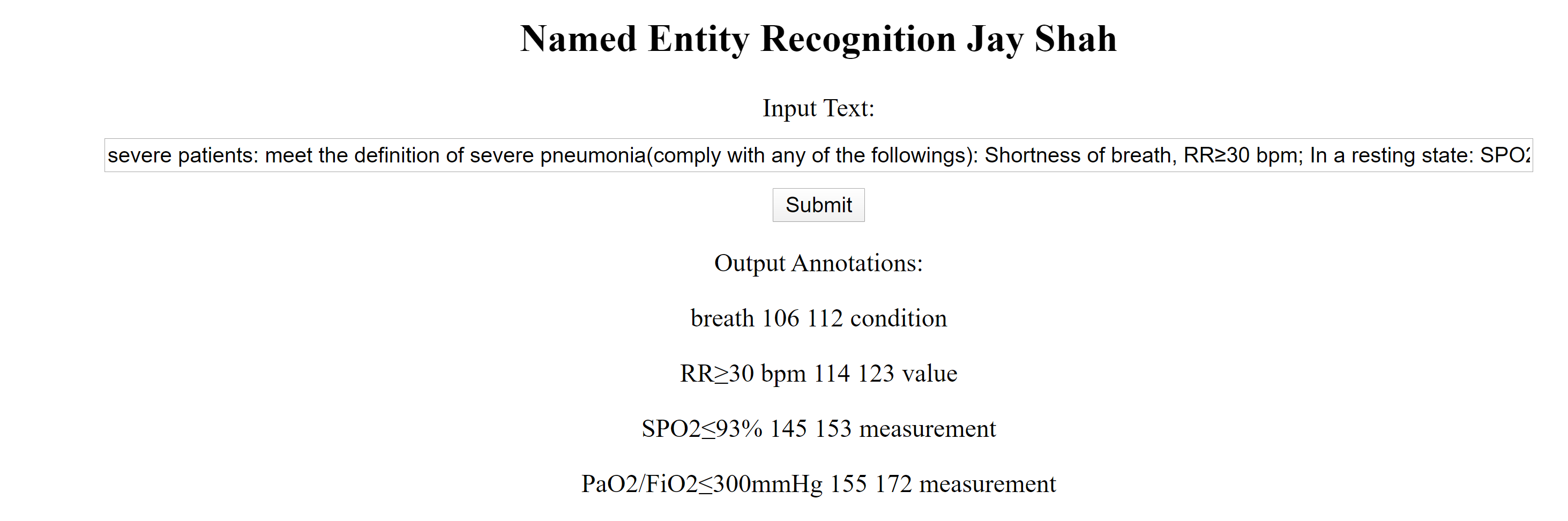
Task 1: Here is the link for the deployed Named Entity Recognition Model

<http://medicalner-env.eba-uhazeecp.us-east-1.elasticbeanstalk.com/>

Here is a sample output:



I have attached the codes for training the model and developing the UI.

Task 2:

The following is an Active Learning framework Jason can use for the annotation procedure of medical text:

(1) *Initial model generation:* In the beginning, a small number of annotated samples can be queried for annotation to build an initial model. We can use longest sentence sampling strategy because it can induce a better initial model or a starting point in the learning curve than random sampling for each querying method.

(2) *Querying:* The unannotated sentences can then be ranked based on the querying algorithm. We can use updated the Conditional Random Fields (CRF) model for ranking (e.g. uncertainty sampling). A number of top ranked sentences can then be selected for annotation, and then added to the annotated set. The batch sizes (the size of top ranked sentences selected for annotation) of each iteration can be in the order of 2(i + 2), where *i* is the number of iterations. This is one of the standard ways to select batch size for the active learning experiment and has been used in an active learning challenge.

(3) *Training:* The CRF model can then be retrained on the updated annotated set.

(4) *Iteration:* Repeat Steps (2) and (3) until the stop criterion is met. In this study, the annotation stops when all sentences in the pool of unlabeled set are queried.

Multiple measurements should be stored during the active learning process for evaluation, such as model quality in F-measure, number of words in the annotated set, and number of entities in the annotated set.

Uncertainty-based querying algorithms

The assumption here is that the most uncertain sentences are most informative because identification of their uncertain labels could gain the most utility for the supervised NER learning. Only the *N*-best annotations should be considered since the size of the possible sequence labels grows exponentially as the length of a sentence increases. We can also extend the *N*-best annotations to cover most of the highly possible annotations. The following three can be used to calculate the uncertainty of a sentence:

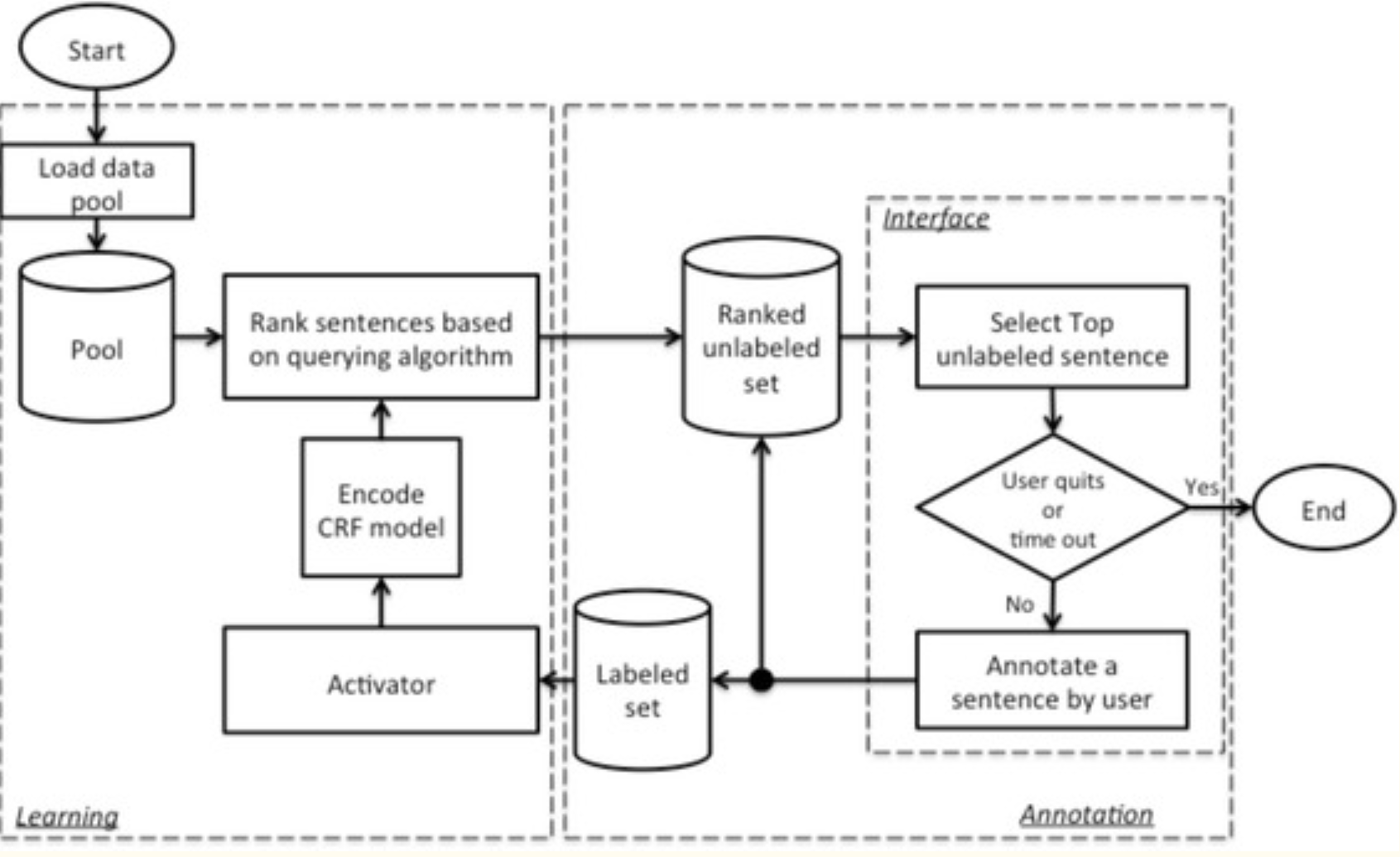
(1) Least Confidence (*LC*): to take the uncertainty from the best possible annotations based on the posterior probability output from CRF. The uncertainty of a sentence *x* is equal to 1 − *P*(*y*∗|*x*), where *y*∗ is the most likely sequence label for the sentence *x*.

(2) *Margin*: to take the uncertainty from the best two possible sequence labels. The uncertainty of a sentence *x* is equal to *P*(*y*∗|*x*) − *P*(*y*∗∗|*x*), where *y*∗ and *y*∗∗ are the most likely and second most likely sequence labels, respectively, for the sentence *x*. The lower margin between the two probabilities represents higher uncertainty.

(3) *N*-*best sequence entropy*: to take the entropy of the probability distribution over *N*-best sequence labels predicted by the CRF model.

An Active learning approach Jason can use for the annotation procedure on a large pool of text data:

Once the system starts, the pool of unlabeled data must loaded. At the initial iteration or before the CRF model is generated, all sentences are randomly ranked. The top sentence in the ranked unlabeled set is queried and displayed on the interface. The user then can highlight clinical entities in the sentence via the labeling function on the interface (annotation process). When the user submits the annotated sentence, the labeled set and the unlabeled set are updated and the learning process is activated based on activation criteria. When the learning process is complete, the ranked unlabeled set is updated while the next sentence is available for annotation.



One sub-flow runs back to the ranked unlabeled set and interface. Therefore, the user can immediately read the next sentence on the interface right after the annotation for one sentence is submitted. The other sub-flow adds the newly annotated sentence to the labeled set, which is then pushed to the learning process. A new learning process will be activated if the encoding or querying process is not busy and the number of newly annotated sentences is greater or equal to a threshold, which is for the update frequency control. When the learning process is activated, it runs in parallel with the annotation process and it updates the ranked unlabeled set whenever the new rankings are generated. This design allows a user to continuously annotate the top unlabeled sentence from the ranked list, which is generated in either the current or previous learning process. The program can be stopped when the user either clicks the quit button or a pre-set cutoff time runs out.

The learning process includes CRF model encoding based on the current labeled set and sentence ranking by the querying engine.  Sentence ranking consists of two steps: 1) CRF model decoding, which is to make predictions for each unlabeled sentence based on the current model; and 2) ranking sentences by the querying algorithm, which considers both the probabilistic prediction of each sentence from the first step, and other information about the unlabeled sentences.

Querying Method:

Jason can use Clustering And Uncertainty Sampling Engine (CAUSE) that combines clustering technique and uncertainty sampling to query both informative and representative sentences. This method guarantees that the top ranked sentences in a batch are from different clusters and thus dissimilar with each other.

**Input**

1. Clustering results of sentences; (2) Uncertainty scores of sentences; (3) Batch size (x)

**Steps**

(1) Cluster ranking: score each cluster based on the uncertainty scores of sentences and select the top x cluster(s) based on the cluster scores, where x is the batch size; (e.g. the score of a cluster could be the average uncertainty score of sentences in this cluster.)

(2) Representative sampling: in each selected cluster, find a sentence with the highest uncertainty score as the cluster representative.

**Output**

x cluster representative sentences in the order of their cluster ranking

**Sentence clustering with topic modeling**

Clustering is a required pre-processing step in CAUSE for the pool of data to be queried. The clustering process consists of Latent Dirichlet Allocation (LDA), a topic modeling technique, for feature generation, and affinity propagation (AP) for clustering. In this clinical concept extraction task, we need to group semantically similar sentences together. LDA can be applied to extract the hidden semantic topics in the corpus of clinical notes. Using document-level samples for topic modeling could generate more meaningful topics than sentences. Given the K estimated topics, the LDA inference process can be performed to assign probabilistic values of topics for every sentence. Eventually, each sentence can be coded in a K dimensional vector with a probability at each of the K topics as value. Similarity can be calculated between every sentence pair. Affinity propagation can be used that takes the M x M pair-wise similarity matrix as the input and outputs the clustering result for the M sentences.

A score based on Average uncertainty cluster sampling (AUCS) can be applied to every cluster (assigning the cluster the average uncertainty score from all the sentences in the cluster).

**Generating Annotations:**

From the top ranked cluster, we select the sentence that has the highest uncertainty score as the representative of the cluster. We also find the representative sentences from the second ranked cluster, third ranked cluster, and so on. We keep sampling until the batch is filled up with representatives. The ranking of the representatives follows the ranking of their clusters. We assume that the number of clusters is greater than or equal to the batch size so that the batch cannot contain more than one sentence from a cluster.

E.g for a cluster representative where fever is annotated as a condition; when the NER model is trained on this annotated cluster representative, the model could identify other conditions (e.g. “breath”, “diarrhea” etc.) from additional sentences in the same cluster based on their similar context as the cluster representative.

Hence, Jason can Use CAUSE algorithm along with CRF for uncertainty sampling to implement an active learning method for the annotation procedure.