Active learning: a step towards automating medical concept extraction

Mahnoosh Kholghi1,2, Laurianne Sitbon1, Guido Zuccon1, Anthony Nguyen2

ABSTRACT

Objective This paper presents an automatic, active learning-based system for the extraction of medical concepts from clinical free-text reports. Specifically, (1) the contribution of active learning in reducing the annotation effort and (2) the robustness of incremental active learning framework across different selection criteria and data sets are determined.

Materials and methods The comparative performance of an active learning framework and a fully supervised approach were investigated to study how active learning reduces the annotation effort while achieving the same effectiveness as a supervised approach. Conditional random fields as the supervised method, and least confidence and information density as 2 selection criteria for active learning framework were used. The effect of incremental learning vs standard learning on the robustness of the models within the active learning framework with different selection criteria was also investigated. The following 2 clinical data sets were used for evaluation: the Informatics for Integrating Biology and the Bedside/Veteran Affairs (i2b2/VA) 2010 natural language processing challenge and the Shared Annotated Resources/Conference and Labs of the Evaluation Forum (StARE/CLEF) 2013 eHealth Evaluation Lab.

Results The annotation effort saved by active learning to achieve the same effectiveness as supervised learning is up to 77%, 57%, and 46% of the total number of sequences, tokens, and concepts, respectively. Compared with the random sampling baseline, the saving is at least doubled.

Conclusion Incremental active learning is a promising approach for building effective and robust medical concept extraction models while significantly reducing the burden of manual annotation.

BACKGROUND AND SIGNIFICANCE

The goal of active learning (AL) is to maximize the effectiveness of the learning model while minimizing the number of annotated samples required. The main challenge is to identify the informative samples that guarantee the learning of such a model.6 Settles and Craven13 reported an extensive empirical evaluation of a number of AL selection criteria using different corpora for sequence labeling tasks. Information density (ID), sequence vote entropy, and least confidence (LC) were found to outperform the state-of-the-art in AL for sequence labeling in most corpora.

While the effectiveness of AL methods has been conclusively proven and demonstrated in many domains for tasks such as text classification, information extraction (IE), and speech recognition,1 as Ohno-Machado et al17 highlighted, there are limited explorations of AL techniques in clinical and biomedical natural language processing (NLP) tasks. There are examples of using AL for classifying medical concepts according to their assertions14,15 and coreference resolution,16 both of which are 2 important elements of any clinical IE system. For assertion classification, Chen et al6 introduced a "model change" sampling-based algorithm which controls the changes of certain values from different models during the AL process. They found it performing better in terms of effectiveness and annotation rate than uncertainty sampling and ID selection criteria.

In the medical IE context, AL has also been used for de-identifying Swedish clinical records.17 The most uncertain and the most certain sampling strategies were evaluated using the highest and lowest entropy, respectively. The evaluation showed that these methods outperformed other benchmarks, including a random sampling (RS) baseline. Figueroa et al18 analyzed the performance of distance-based and diversity-based algorithms as 2 AL methods, in addition to a combination of both, for the classification of smoking and depression status. The performance of the proposed methods in terms of accuracy and annotation rate was found to be strongly dependent on the data set diversity and uncertainty.

OBJECTIVE

The widespread use of e-Health technologies, in particular of electronic health records, offers many opportunities for data analysis.1 The extraction of structured data from unstructured, free-text health documents (eg, clinical narratives) is essential to support such analysis, in particular for applications such as retrieving, reasoning, and reporting, eg, for cancer notification and monitoring from pathology reports.2,3 This extraction process commonly consists of a concept extraction stage, ie, the identification of short-term sequences (eg, named entities, phrases, or others) from unstructured text written in natural language. These terms express meaningful concepts within a given domain, eg, medical problems, tests, and treatments (figure 1) from the i2b2/VA 2010 data set.4

This IE process, however, is not straightforward. Challenges include the identification of concept instances that are referred to in ways not captured within current lexical resources and the presence of ambiguity, polysemy, synonymy (including acronyms), and word order variations. In addition to these challenges, the information presented in clinical narratives is often unstructured, ungrammatical, and fragmented, and language usage greatly differs from that of general free-text. Because of this, mainstream NLP technologies and systems often cannot be directly used in the health domain.5 In fact, high-quality manual annotations of large corpora are necessary for building...
robust statistical supervised machine learning (ML) classifiers. Obtaining these annotations is costly because it requires extensive involvement of domain experts (eg, pathologists or experienced clinical coders to annotate pathology reports) and linguists. AL has been proposed as a feasible alternative to standard supervised ML approaches to reduce annotation costs.6 We embrace this proposal, and in this article we investigate the use of active learning for medical concept extraction.

AL methods use supervised ML algorithms in an iterative process, where at each iteration, samples from the data set are automatically selected based on their “informativeness” to be annotated by an expert. This is effectively a human-in-the-loop process that allows to drastically reduce human involvement compared with the large amount of annotated data required upfront by standard supervised ML systems. Despite this interesting property, AL has not been fully explored for clinical IE, in particular for medical concept extraction.7

The aim of this article is to extensively investigate the comparative performance of a fully supervised ML approach and an AL counterpart for the task of extracting medical concepts related to problems, treatments (i2b2/VA 2010 task1), and disorder mentions (ShARe/CLEF 2013, task 1 8). Specifically, the following hypotheses are validated:

1. AL achieves the same effectiveness as a supervised approach using much less annotated data, and (2) incremental AL results in more robust learnt models than standard AL regardless of selection criteria. Preliminary investigations suggest that a selected feature set and incremental learning could increase robustness.9 However, it has not been examined across different settings and selection criteria.

To investigate our hypotheses, we use conditional random fields (CRFs)10 as the supervised IE component. CRFs are a proven state-of-the-art technique for the task at hand.4,11 Two AL selection criteria are investigated, namely uncertainty sampling-based12 and information density (ID).13 Annotated data from the i2b2/VA 2010 Clinical NLP Challenge and the ShRe/CLEF 2013 eHealth Evaluation Lab (task 1) are used for training the models and evaluating their performance.

FEATURES FOR MACHINE LEARNING FRAMEWORKS

The following groups of features are used for both the fully supervised and the AL methods:

- Linguistic and orthographical features (eg, regular expression patterns and part-of-speech tags)
- Lexical and morphological features (eg, suffixes/prefixes and character n-gram)
- Contextual features (eg, window of k words)
- Semantic features (eg, SNOMED CT and UMLS semantic groups from Medtex,20 a medical NLP toolkit)

Fully supervised approach
We use linear-chain CRFs as a supervised IE algorithm and as a base algorithm in the AL framework. Let \( x = (x_1, \ldots, x_n) \) and \( y = (y_1, \ldots, y_n) \) be an observation sequence and its corresponding label sequence, respectively. For example, sentence 2 in figure 2 would correspond to \( x = (1, 2, 1, 30) \) and \( y = (0, 0, 0, 0) \) the current treatment, treatment. The posterior probability of \( y \) given \( x \) is described by the linear-chain CRFs model with a set of parameters \( \theta \):

\[
P_y (y | x) = \frac{1}{Z(x)} \exp \left( \sum_{j=1}^{n} \sum_{i=1}^{m} \lambda_j f_j(y_{j-1}, y_j, x_j) \right)
\]

where \( Z(x) \) is the normalization factor. Each \( f_j \) is the transition feature function between label states \( j-1 \) and \( j \) on the sequence \( x \) at position \( i \). The \( \theta = (\lambda_1, \ldots, \lambda_m) \) parameters represent the corresponding feature weights.

An important parameter when training CRFs is the Gaussian prior variance or regularization parameter. In our prior investigation,9 we found that the optimal value for clinical data was 1.

Active learning
The AL framework described by the algorithm in figure 3 is characterized by 6 main elements:

1. The initial labeled set is usually a very small, randomly selected portion (eg, <1%) of the whole data set;
2. A batch of instances is selected at each iteration of the AL algorithm;
3. The actual instances, in our case, correspond to a complete sequence or sentence, thus yielding a sequence-based AL method (rather than a token-based or document-based);

<table>
<thead>
<tr>
<th>Line No.</th>
<th>Input Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>Per Cardiology, it was felt that the patient had no clinical evidence of any CHF by physical examination or x-ray; however, the patient did improve subjectively on Lasix, prednisone as well as doxycycline.</td>
</tr>
<tr>
<td>62</td>
<td>It was recommended to continue current treatment.</td>
</tr>
</tbody>
</table>

Concepts ("concept text" start(\#Line\#Token\#) end(\#Line\#Token\#) || t="concept type")
- c="any cbf" 61:14 61:15 || t="problem"
- c="physical examination" 61:17 61:18 || t="label"
- c="x-ray" 61:20 61:20 || t="label"
- c="Lasix" 61:30 61:30 || t="label"
- c="Prednisone" 61:32 61:32 || t="label"
- c="doxycycline" 61:36 61:36 || t="label"
- c="current treatment" 62:5 62:6 || t="label"

Figure 1: An example of input text and associated concepts from the i2b2/VA 2010 data set.
4. The stopping criterion identifies the condition to be met for terminating the AL process. As we aim to study how AL can contribute towards reducing the annotation effort compared to a supervised approach, we use supervised effectiveness as our target performance.

5. The selection criterion $\phi(u_i, \Theta)$ estimates the informativeness of an unlabeled instance $u_i \in \mathcal{U}$ based on the model $\Theta$. The confidence of the latest model is used in predicting the label $y$ of a sequence $x$.

$$\phi_{\text{LC}}(x, \Theta) = 1 - P_{\text{HC}}(y | x)$$

(2)

The confidence of the CRFs model is estimated using the posterior probability described in Equation (1), and $y$ is the predicted label sequence obtained with the Viterbi algorithm.

6. The model can either be achieved by fully retraining the model using all labeled data at each step or incrementally by updating the model learnt in the preceding loop with new labeled instances.

The most critical part when designing an AL strategy is how to estimate the informativeness of each instance, ie, selecting an effective selection criterion.

### Uncertainty sampling

One of the most common selection criteria is uncertainty sampling.\(^{12}\) According to this paradigm, instances with the highest uncertainty are selected for labeling and inclusion in the labeled set used for training the model in the following iteration. We propose to use LC as one of our selection criteria. LC uses the confidence of the latest model in predicting the label $y$ of a sequence $x$:

$$\phi_{\text{LC}}(x, \Theta) = 1 - P_{\text{HC}}(y | x)$$

(2)

Besides a fully supervised approach, common baselines for analyzing the benefits of the AL framework are RS and Longest Sequence (LS). Both baselines follow the steps of figure 3, except for RS randomly selects instances and LS chooses instances with the longest length in each batch.

### Incremental learning

In the standard AL approach, at each iteration, instances are selected from $\mathcal{U}$ according to a selection criterion, manually labeled, and then added to the labeled set $\mathcal{L}$. A new model is built on $\mathcal{L}$ independently of models built in previous iterations. In other words, a supervised model is built from $\mathcal{L}$ (step 1 in figure 3) at each iteration.

The alternative is to use an incremental approach. Here all weights and values of the model learnt in previous iteration are maintained, and they are updated in the current iteration using the corresponding labeled set. It makes the training of models faster than the standard setting, leading to considerable reduction in processing time across the whole AL loop. Through the comparison of the standard and the incremental AL approaches, we aim to unveil the effect that maintaining models’ weights and values across iterations has on the robustness of the models themselves within the considered clinical IE tasks.

$$\text{sim}_{\text{cos}} \left( \overrightarrow{x}, \overleftarrow{x}^{(u)} \right) = \frac{\overrightarrow{x} \cdot \overleftarrow{x}^{(u)}}{\|\overrightarrow{x}\| \|\overleftarrow{x}^{(u)}\|}$$

(4)

Where $\overrightarrow{x}$ refers to the feature vector of instance $x$, $R_{\text{representative}}(x)$ is calculated as the mean of the similarities between $x$ and all sequences across the unlabeled set:

$$R_{\text{representative}}(x) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{sim}(x, x^{(u)})$$

(5)
mental AL, including LC and ID, are based on the MALLET toolkit.22 Our implementations of CRFs for supervised learning, RS, and incremental active learning framework; InALCE-Tun, tuned incremental active learning framework; Sup, supervised approach; Sup-Tun, tuned supervised approach.

## Data sets

For generating learning curves, we used the annotated train and test sets from the concept extraction task in the i2b2/VA 2010 NLP challenge and ShARe/CLEF 2013 eHealth Evaluation Lab (task 1).11 One of the i2b2/VA 2010 tasks was to extract medical problems, tests, and treatments from clinical reports. The training and testing sets included 349 and 477 reports, respectively.

Task 1 of the ShARe/CLEF 2013 eHealth Evaluation Lab was to extract and identify disorder mentions from clinical notes. The data set consisted of 200 training and 100 testing clinical notes.

## Experimental settings

### Data sets

For generating learning curves, we used the annotated train and test sets from the concept extraction task in the i2b2/VA 2010 NLP challenge and ShARe/CLEF 2013 eHealth Evaluation Lab (task 1).11 One of the i2b2/VA 2010 tasks was to extract medical problems, tests, and treatments from clinical reports. The training and testing sets included 349 and 477 reports, respectively.

Task 1 of the ShARe/CLEF 2013 eHealth Evaluation Lab was to extract and identify disorder mentions from clinical notes. The data set consisted of 200 training and 100 testing clinical notes.

### Active learning selection criteria

**Table 1: Annotation rate for ALCE and InALCE settings using least confidence as selection criterion**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Selection Criteria</th>
<th>ALCE</th>
<th>InALCE</th>
<th>InALCE-Tun</th>
</tr>
</thead>
<tbody>
<tr>
<td>i2b2 2010</td>
<td>Sup (F1 = 0.8018)</td>
<td>30</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Sup-Tun (F1 = 0.8212)</td>
<td>NA</td>
<td>NA</td>
<td>23</td>
</tr>
<tr>
<td>CLEF 2013</td>
<td>Sup (F1 = 0.6579)</td>
<td>44</td>
<td>35.5</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Sup-Tun (F1 = 0.6689)</td>
<td>NA</td>
<td>NA</td>
<td>30</td>
</tr>
</tbody>
</table>

Abbreviations: NA, not available; SAR, sequences annotation rate; ALCE, nontuned standard active learning framework; InALCE, nontuned incremental active learning framework; InALCE-Tun, tuned incremental active learning framework; Sup, supervised approach; Sup-Tun, tuned supervised approach.

### Results

In our evaluation, concept extraction effectiveness is measured by Precision, Recall, and F1-measure. Learning curves highlight the interaction between model effectiveness and the required annotation effort.

The lower the AR, the less annotation effort is required. Here we assume that every instance is considered as having the same annotation effort (uniform annotation effort). However, in reality, sentences could be short or long. Longer sentences with more entities would require more time for annotation. But shorter sentences without any entities could be skipped by annotators quickly, thus taking little time for annotating them. While this setting may not be fully representative of real-world use-cases, actual annotation costs are not available for the considered data sets, and the literature lacks specific studies that consider annotation cost models for medical concept extraction. Modeling annotation cost is outside of the scope of this paper. However, our evaluation provides an indication of the reduction in annotation effort that AL could contribute.

In addition, we perform 10-fold cross validation experiments on the training data to analyze the robustness of AL and RS within the incremental settings in both data sets and across different selection criteria.

### Incremental vs standard active learning

As discussed in prior work, incremental AL framework with tuned Gaussian prior variance for CRFs (InALCE-Tun) leads to models that are more stable across batches and also achieves higher effectiveness compared with the nontuned standard AL framework (ALCE) and nontuned incremental AL framework (InALCE). We now investigate the annotation rate across these frameworks.

The results reported in table 1 suggest that InALCE reaches the same performance as the supervised approach in both data sets but with a faster learning rate than ALCE. Furthermore, by tuning the CRFs parameter in the incremental AL framework (InALCE-Tun), not only the supervised performance increases but also the annotation rate decreases, which means lower annotation effort is required to reach the target performance. We then retain InALCE-Tun for the next experimental settings.

### Active learning selection criteria

So far, we have only considered the LC selection criterion. In this section, we compare its performance with that of the ID selection criterion, with RS and LS as baselines.

Figure 4 shows the performance of these with InALCE-Tun on i2b2/VA 2010 and ShARe/CLEF 2013, highlighting that the use of a different selection criterion does not have an impact on the robustness gained by using InALCE-Tun. In both data sets, we observe that LC and ID reach the target performance quicker than the RS and LS baselines. Furthermore, it can be noticed that these 2 criteria always outperform the baselines, and that there is no noticeable difference between LC and ID. The annotation rates from table 2 show that LC requires less annotation effort to reach the target performance. ID is computationally costly compared to LC. Depending on the size of the data set, it may require a large amount of similarity calculations for all instances in the unlabeled set, which could be precomputed before running AL. While this would not be a problem in real-time AL systems, the results still suggest choosing LC a
priori as a selection strategy because it always achieves a slightly lower AR.

Effectiveness of AL beyond the target performance

Results in Table 3 consider the batch in which AL approaches achieve the highest F1-measure, which is generally beyond the target (supervised) performance (where the AL learning curve and the target intersect). These results demonstrate that LC can outperform the supervised method using 44% of the whole training data for i2b2/VA 2010 and using 48% of the whole training data for the ShARe/CLEF 2013 task. The highest performance rates reported in Table 3 suggest that LC requires less training data than ID to achieve the highest
The same finding is however not confirmed by the ShARe/CLEF 2013 results, for which no difference is observed.

**Robustness Analysis**

To analyze the robustness of AL models in the incremental setting (InALCE-Tun), we perform a 10-fold cross validation on both i2b2/VA 2010 and ShARe/CLEF 2013 training sets. In these experiments, the training set is split in 10 random sets; 9 are used as labeled training data and 1 as testing data. This process is then iterated by varying which fold is used for testing.

AL is applied throughout the training data and the performance of the learnt model at each batch is averaged across the testing folds. Figure 5 shows the learning curves of InALCE-Tun with different selection criteria across the 2 data sets; these are obtained considering at each batch the mean performance of the learnt model and its variance across the cross-validation folds. AL models built across batches in the cross validation setting are robust, as they show small variance across folds.

**DISCUSSION**

The results of our empirical investigation demonstrate that AL strategies clearly have a role to play in reducing the burden of annotation for high-quality medical concept extraction. We found that the AL selection criteria always perform better than the baselines (RS and LS) in terms of effectiveness and annotation rate. While RS does not reach the target supervised learning performance much before having trained on all batches, it is interesting to note that it does come within 10% of the target performance after only 26 (5273 instances) out of...
153 and 30 (912 instances) out of 91 iterations in i2b2/VA 2010 and ShARe/CLEF 2013, respectively. This suggests that the data sets used in our study present patterns that are often repeated between instances. It is also important, however, to remember that the training instances are sentences; thus, on average, after 26 to 30 iterations there could be nearly 10 sentences from each document in the training set. Repeated patterns, especially repeated concepts themselves, are to be expected within patient records, and therefore it should be expected that the entire training set presents redundancies.

Furthermore, it only takes AL 7 iterations (1400 instances) in i2b2/VA 2010 and 6 iterations (200 instances) in ShARe/CLEF 2013 to reach an F1-measure value within 10% of that reached by the supervised CRFs model. This is a very important result in contexts where one may be able to balance the effectiveness against annotation effort, eg, when performing concept extraction to build a knowledge base.

Our comparison of various AL methods shows that, in the considered task, the LC selection criterion outperforms the ID criterion. This suggests that the probability distribution of the data in clinical narratives is representative of the characteristics of the data. In addition, we verified that an incremental AL approach (InALCE-Tun) provides higher robustness than the standard AL approach (ALCE) and it requires less labeled data to reach the target performance.

Despite ID being regarded as the state-of-the-art in AL, this criterion was not able to outperform the LC approach within the medical concept extraction task considered in this paper. We speculate that this is because of the high similarity found in clinical narratives. The ID method tends to select samples from dense regions of the data to avoid selecting outliers; thus, samples that are more similar to other samples in the data set are more likely to be selected.

Figure 6 shows the distribution of full duplicates in the i2b2/VA 2010 data set based on the sequence length. We found that 35% of the total number of sequences in the data set was exactly replicated. As shown in figure 6, 71% of full duplicates are sequences with lengths between 1 and 3, which are less likely to contain the target concepts. To investigate if there is any effect of full duplicates on AL performance, the experiments were replicated with a precondition in the steps of the AL process, which prevented the algorithm from selecting full duplicates. The annotation rates are reported in table 4 against the same data set and target performance as in table 2.

Results show that ignoring full duplicates have no significant effect on LC and ID performance in terms of annotation rates. It is not surprising as LC and ID are likely to select almost the same informative instances as in previous settings. Indeed, LC selects sequences containing labels where the model has low confidence, and an already seen sequence would immediately have high probabilities associated to its labels.

Although full duplicates would potentially have been selected by ID due to their high similarity, the probability component in the ID function (Equation (3)) avoids choosing those instances as the model is confident enough about their labels.

The differences for RS results from table 2 are due to the fact that the probability of selecting full duplicates for RS is now zero, which means useful samples are more likely to be selected at each iteration.

We also found that a fully supervised model trained on a duplicate-free version of the data set yielded almost the same performance (F1 measure = 0.8224) as on the full training set (F1 measure = 0.8212), which suggests that CRFs models do not make use of repeated sequences.

We analyzed the errors performed in the considered concept extraction task by considering the confusion matrices obtained from the classification results (reported in supplementary appendix A), and we observed that in both data sets the largest amount of errors is found to be the misclassification of a target entity into a nontarget entity (eg, “problem,” “test,” or “treatment” entities misclassified as “others” in i2b2/VA 2010). In this data set, the target problem presents the least classification errors; this is consistent across all approaches. The error analysis using confusion matrices also suggested that the AL model learnt up to the target performance is similar (in terms of classification errors) to the model learnt by the supervised approach on the whole data set, suggesting AL does not overfit the data.

In summary, we can conclude that (1) AL can reach at least the same effectiveness of supervised learning for medical concept extraction while using less training data, and (2) the IE models learnt by AL are not overfitted as they appear to lead to similar errors to those from the models learnt by the fully supervised approach.

CONCLUSION

This paper presented a simulated study of AL for medical concept extraction. We have empirically demonstrated that AL can be highly effective for reducing the effort of manual annotation while building reliable models. We demonstrated this by comparing the effectiveness of fully supervised CRFs against 2 AL approaches and 2 baselines. The evaluation based on the i2b2/VA 2010 NLP challenge and the ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) data sets showed that AL (specifically when using the LC selection criterion) achieves the same effectiveness of supervised learning using 54% (i2b2/VA 2010) and 76% (ShARe/CLEF 2013) of the total number of concepts in training data. We also showed that incremental learning leads to more reliable models within the AL framework.

While this research contributes a very important first step in introducing AL for medical concept extraction, further work is required to examine other selection criteria and develop a cost model for evaluation of AL.

CONTRIBUTORS

This article is a part of MK’s PhD thesis. She developed the ideas and wrote all the coding. LS, GZ, and AN supervised, analyzed, and discussed the results of MK’s work. All authors have contributed in writing and revising the manuscript.

COMPETING INTERESTS

None.
ACKNOWLEDGEMENTS
We would like to thank the anonymous reviewers for their helpful feedback.

SUPPLEMENTARY MATERIAL
Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

AUTHOR AFFILIATIONS

1Science and Engineering Faculty, Queensland University of Technology, Brisbane 4000, Queensland, Australia.
2The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Queensland, Australia.