ACTIVE LEARNING: A STEP TOWARDS AUTOMATING MEDICAL CONCEPT EXTRACTION

Corresponding Author: Mahnoosh Kholghi[1,2], Queensland University of Technology, Gardens Point Campus, 2 George St, Brisbane, Queensland 4000, Australia, email: m1.kholghi@qut.edu.au, phone number: +61 403320109

Co-authors: Laurianne Sitbon[1], Guido Zuccon[1], Anthony Nguyen[2]

[1]Science and Engineering Faculty, Queensland University of Technology, Brisbane 4000, Queensland, Australia.

[2]The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Queensland, Australia

Running Title: Active Learning For Medical Concept Extraction

Key words: Medical Concept Extraction, Clinical Free Text, Active Learning, Conditional Random Fields, Robustness Analysis.

Word Count: 4000

PRE-PRINT VERSION

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 datasets is 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 two 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. Two clinical datasets were used for evaluation: the i2b2/VA 2010 NLP challenge and the ShARe/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 to the Random sampling baseline, the saving is at least doubled.

Discussion

Incremental active learning guarantees robustness across all selection criteria and datasets. The reduction of annotation effort is always above random sampling and longest sequence baselines.

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.

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 (e.g., clinical narratives) is essential to support such analysis, in particular for applications such as retrieving, reasoning, and reporting, e.g., for cancer notification and monitoring from pathology reports [2, 3]. This extraction process commonly consists of a concept extraction stage, i.e. the identification of short-term sequences (e.g., named entities, phrases or others) from unstructured text written in natural language. These terms express meaningful concepts within a given domain (e.g., medical problems, tests, and treatments (Figure 1) from the i2b2/VA 2010 dataset [4]).

This information extraction (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 natural language processing (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 (e.g., pathologists or experienced clinical coders to annotate pathology reports) and linguists.

Active learning (AL) has been proposed as a feasible alternative to standard supervised machine learning 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 dataset 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 to the large amount of annotated data required upfront by standard supervised ML systems. Despite this interesting property, active learning has not been fully explored for clinical information extraction, in particular for medical concept extraction [7].

The aim of this article is to extensively investigate the comparative performance of a fully supervised machine learning approach and an active learning counterpart for the task of extracting medical concepts related to problems, tests and treatments (i2b2/VA 2010 task [4]) and disorder mentions (ShARe/CLEF 2013 (task 1) [8]). Specifically, the following hypotheses are validated: (1) active learning achieves the same effectiveness as a supervised approach using much less annotated data, and (2) incremental active learning results in more robust learnt models than standard active learning regardless of selection criterion. Preliminary investigations suggested that 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 information extraction component: CRFs is a proven state-of-the-art technique for the task at hand [4, 11]. Two active learning selection criteria are investigated, namely uncertainty sampling-based [12] and information density [13]. Annotated data from the i2b2/VA 2010

Clinical NLP Challenge and the ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) are used for training the models and evaluating their performance.

BACKGROUND AND SIGNIFICANCE

The goal of active learning 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 to learn such a model [6].

Settles and Craven [13] reported an extensive empirical evaluation of a number of active learning selection criteria using different corpora for sequence labeling tasks. Information density, sequence vote entropy, and least confidence were found to outperform the state of the art in active learning for sequence labeling in most corpora.

While the effectiveness of active learning methods has been conclusively proven and demonstrated in many domains for tasks such as text classification, information extraction and speech recognition [6], as Ohno-Machado, et al. [1] highlighted, there are limited explorations of AL techniques in clinical and biomedical NLP tasks. There are examples of using active learning for classifying medical concepts according to their assertions [14, 15] and co-reference resolution [16], both of which are two important elements of any clinical IE system. For assertion classification, Chen, et al. [14] 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 information density selection criteria.

In the medical information extraction 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 baseline. Figueroa, et al. [18] analyzed the performance of distance-based and diversity-based algorithms, as two active learning 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 dataset diversity and uncertainty.

MATERIALS AND METHODS

Tags for entity representation

Before applying machine learning algorithms to the free-text data, as shown in Figure 1, the annotated data needs to be represented with an appropriate tagging format. We use the "BIO" format, where **B** refers to the "beginning" of an entity, **I** refers to the "inside" of an entity, and **O** refers to the "outside" of an entity [19]. Figure 2 shows the BIO tag representation for the sentences from Figure 1.

Features for machine learning frameworks

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

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

Fully supervised approach

We use linear-chain CRFs as supervised information extraction algorithm and as base algorithm in the AL framework. Let $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$ be an observation sequence and its corresponding label sequence, respectively. For example, sentence 2 in Figure 2 would correspond to $\vec{x} = (\text{It}, \text{was}, \text{recommend}, \text{to}, \text{continue}, \text{current}, \text{treatment})$ and $\vec{y} = (O, O, O, O, O, B-\text{treatment}, I-\text{treatment})$. The posterior probability of $\vec{y}$ given $\vec{x}$ is described by the linear-chain CRFs model with a set of parameters θ:

$$P_\theta(\vec{y}|\vec{x}) = \frac{1}{Z_\theta(\vec{x})} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_i)\right) \quad (1)$$

where $Z_\theta(\vec{x})$ is the normalization factor. Each f_j is the transition feature function between label states $i-1$ and i 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 active learning framework described by the algorithm in Figure 3 is characterized by six main elements:

1. The *initial labeled set* is usually a very small, randomly selected, portion (e.g., less than 1%) of the whole dataset;
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);

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* $\varphi(u_i, \Theta)$ estimates the informativeness of an unlabeled instance $u_i \in \mathcal{U}$ based on the model Θ.
6. Training 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, i.e. 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 in the following iteration. We propose to use *Least Confidence* (LC) as one of our selection criteria. LC uses the confidence of the latest model Θ in predicting the label $\vec{y}$ of a sequence $\vec{x}$ [21]:

$$\varphi_{LC}(\vec{x}, \Theta) = 1 - P_\theta(\vec{y}^*|\vec{x}) \quad (2)$$

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

Information density

Information Density (ID) is an alternative selection criterion for AL [13]. The intuition behind ID is that the selection of instances that are both informative and representative lead to a better coverage of the dataset characteristics. By considering also the representativeness of instances, along with their informativeness, outliers are less likely to be selected by the AL process. ID is computed according to:

$$\varphi_{\mathrm{ID}}(\vec{x}, \Theta) = \varphi_{\mathrm{informative}}(\vec{x}, \Theta) \,.\, \mathcal{R}_{\mathrm{representative}}(\vec{x}) \tag{3}$$

Where $\mathcal{R}_{\mathrm{representative}}(\vec{x})$ corresponds to the representativeness of instances. In this study, we use the least confidence (Equation (2)) to measure the informativeness of instances ($\varphi_{\mathrm{informative}}(\vec{x}, \Theta)$). The average similarity between instance $\vec{x}$ and all other sequences in the set of unlabeled instances $\mathcal{U}$ indicates how representative the instance $\vec{x}$ is: the higher the similarity, the more representative the instance is. Similarity between $\vec{x}$ and another sequence is measured according to the cosine distance:

$$sim_{cos}\left(\vec{x}, \vec{x}^{(u)}\right) = \frac{\vec{\mathbf{x}} . \vec{\mathbf{x}}^{(\mathrm{u})}}{\|\vec{\mathbf{x}}\| . \|\vec{\mathbf{x}}^{(\mathrm{u})}\|} \tag{4}$$

Where $\vec{\mathbf{x}}$ refers to the feature vector of instance $\vec{x}$. $\mathcal{R}_{\mathrm{representative}}(\vec{x})$ is calculated as the mean of the similarities between $\vec{x}$ and all sequences across the unlabeled set:

$$\mathcal{R}_{representative}(\vec{x}) = \frac{1}{\mathcal{U}} \sum_{u=1}^{U} sim(\vec{x}, \vec{x}^{(u)}) \tag{5}$$

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

Incremental learning

In the *standard* active learning 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 information extraction tasks.

Datasets

We used the annotated train and test sets from the concept extraction task in the i2b2/VA 2010 NLP challenge [4] and ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) [11] for generating learning curves. One of the i2b2/VA 2010 tasks was to extract medical problems, tests and treatments from clinical reports. The training and testing sets include 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 dataset consists of 200 training and 100 testing clinical notes.

Experimental settings

Our implementations of CRFs for supervised learning, random sampling, and incremental active learning, including least confidence and information density, are based on the MALLET toolkit [22]. For active learning approaches and random sampling, the initial labeled set is formed by randomly selecting 1% of training data. The batch size (B) is set to 200 for i2b2/VA 2010 and 30 for ShARe/CLEF 2013 across all experiments, leading to a total of 153 and 91 batches, respectively. We simulate the human annotator in the AL process by using the annotations provided in the training portions of the two datasets: these thus are treated as the input the annotator will provide to the AL algorithm.

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 point of intersection between the AL learning curve and the target supervised effectiveness is used to measure how much annotation effort is saved by an AL approach. We analyzed the results by considering the Annotation Rate (AR) for sequences (SAR), tokens (TAR) and concepts (CAR). This can be computed as the number of labeled annotation units (sequences, tokens and concepts) used by AL for reaching this point, over the total number of corresponding labeled annotation units used by the supervised method:

$$\mathrm{AR} = \frac{\#\ labeled\ annotation\ units\ used\ by\ AL}{\#\ total\ labeled\ annotation\ units\ used\ by\ supervised\ method} \quad (6)$$

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 [23], actual annotation costs are not available for the considered datasets and the literature lacks of 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 datasets and across different selection criteria.

RESULTS

Incremental vs. standard active learning

As discussed in prior work [9], incremental active learning 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 to the non-tuned standard active learning framework (ALCE) and non-tuned incremental active learning 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 datasets but with a faster learning rate than ALCE. Furthermore, by tuning the CRFs parameter in incremental active learning framework (InALCE-Tun), not only the supervised performance increased [9], but also annotation rate is decreased which means lower annotation effort required to reach the target performance. We then retain InALCE-Tun for the next experimental settings.

Table 1. Annotation rate for ALCE and InALCE settings using LC as selection criterion.

		SAR(%)		
		ALCE	**InALCE**	**InALCE-Tun**
i2b2 2010	**Sup (F1 = 0.8018)**	30%	23%	15%
	Sup-Tun (F1 = 0.8212)	-	-	23%
CLEF 2013	**Sup (F1 = 0.6579)**	44%	35.5%	19%
	Sup-Tun (F1 = 0.6689)	-	-	30%

Active learning selection criteria

So far we only considered the least confidence (LC) selection criterion. In this section, we compare its performance to that of the information density (ID) selection criterion and with random sampling (RS) and longest sequence (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 criteria does not impact on the robustness gained by using InALCE-Tun. We observe that, in both datasets, LC and ID reach the target performance quicker than the RS and LS baselines. Furthermore, it can be noticed that these two 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.

Information density is computationally costly compared to least confidence. Depending on the size of the dataset, it may require a large amount of similarity calculations for all instances in the unlabeled set, which could be pre-computed before running active learning. While this would not

be a problem in real-time active learning systems, the results still suggest choosing LC a priori as a selection strategy because it always achieves slightly lower AR.

Table 2. Annotation rate for LC, ID and RS.

	SAR(%)		TAR(%)		CAR(%)	
	i2b2 2010	**CLEF 2013**	**i2b2 2010**	**CLEF 2013**	**i2b2 2010**	**CLEF 2013**
Random Sampling	89%	85%	89%	88%	89%	87.5%
Longest Sequence	67%	66%	92%	95%	96%	94%
Information Density	24.5%	30%	45%	59%	54%	77%
Least Confidence	23%	30%	43%	57%	54%	76%

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 performance in i2b2/VA 2010. The same finding is however not confirmed by the ShARe/CLEF 2013 results, for which no difference is observed.

Table 3. The highest performance of active learning methods (P = Precision, R = Recall, F1 = F1-measure).

	i2b2/VA 2010				ShARe/CLEF 2013			
	P	**R**	**F1**	**SAR(%)**	**P**	**R**	**F1**	**SAR(%)**
Sup	0.8378	0.8053	0.8212	-	0.7865	0.5819	0.6689	-
ID	0.8429	0.8114	0.8268	47%	0.7934	0.5953	0.6803	48%
LC	0.8444	0.8112	0.8275	44%	0.7911	0.5965	0.6803	48%

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 ten random sets; nine are used as labeled training data and one as testing data. This process is then iterated by varying which fold is used for testing.

Active learning 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 two datasets: these are obtained considering at each batch the mean performance of the learnt model and its variance across the cross-validation folds. Active learning 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 active learning strategies clearly have a role to play in reducing the burden of annotation for high quality medical concept extraction. We found that the active learning selection criteria always perform better than the baselines (random sampling and longest sequence) in terms of effectiveness and annotation rate. While random sampling 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 datasets 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 (1,400 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, for example when performing concept extraction to build a knowledge base.

Our comparison of various active learning methods showed that, in the considered task, the least confidence selection criterion outperforms the information density 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 information density being regarded as the state-of-the-art in active learning [14, 18], this criterion was not able to outperform the least confidence 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 information density 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 dataset are more likely to be selected.

Figure 6 shows the distribution of full duplicates in i2b2/VA 2010 dataset based on the sequence length. We found that 35% of the total number of sequences in dataset is exactly replicated. As shown in Figure 6, 71% of full duplicates are sequences with lengths between 1-3, 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 pre-condition in the steps of the AL process, which prevents the algorithm from selecting full duplicates. The annotation rates are reported in Table 4, against the same dataset and target performance as in Table 2.

Table 4. Annotation rate for RS, LS, ID, and LC on i2b2/VA 2010 dataset when full duplicates are not allowed to be selected in AL process.

	SAR(%)	TAR(%)	CAR(%)
Random Sampling	58%	75%	82%
Longest Sequence	56%	84%	91%
Information Density	24%	46%	56%
Least Confidence	23%	44%	55%

Results show that ignoring full duplicates had 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 dataset 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 Appendix A), and observed that in both datasets the largest amount of errors is found to be the misclassification of a target entity into a non-target entity (e.g., "*problem*", "*test*" or "*treatment*" entities misclassified as "*others*" in i2b2/VA 2010). In this dataset, 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 dataset, suggesting AL does not over-fit 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 information extraction models learnt by AL are not over-fitted 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 active learning for medical concept extraction. We have empirically demonstrated that active learning 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 two active learning approaches and two baselines. The evaluation based on the i2b2/VA 2010 NLP challenge and the ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) datasets showed that active learning (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 active learning framework.

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

Funding Statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors**.**

Competing Interests Statement

The authors have no competing interests to declare**.**

Contributorship Statement

This article is a part of Mahnoosh Kholghi's PhD thesis. She has developed ideas and done all coding. Laurianne Sitbon, Guido Zuccon and Anthony Nguyen are her supervisors. All results are discussed and analyzed with the supervisors. All authors have contributed in writing and revising the manuscript.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback.

REFERENCES

1. Ohno-Machado L, Nadkarni P, Johnson K. Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature. Journal of the American Medical Informatics Association 2013;**20**(5):805.
2. Nguyen A, Moore J, Lawley M, et al. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. *Health Informatics Conference*. Brisbane, Australia, 2011:117-24.
3. Zuccon G, Wagholikar AS, Nguyen AN, et al. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology. AMIA Summit on Clinical Research Informatics (CRI) 2013;**2013**:300-04.
4. Uzuner Ö, South BR, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 2011;**18**(5):552-56.
5. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association 2011;**18**(5):544-51.
6. Settles B. *Active learning*: Morgan & Claypool Publishers, 2012.
7. Skeppstedt M. Annotating named entities in clinical text by combining pre-annotation and active learning. *Proceedings of the ACL Student Research Workshop*, 2013:74-80.
8. Suominen H, Salanterä S, Velupillai S, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B, eds. Information Access Evaluation. Multilinguality, Multimodality, and Visualization: Springer Berlin Heidelberg, 2013:212-31.

9. Kholghi M, Sitbon L, Zuccon G, et al. Factors influencing robustness and effectiveness of conditional random fields in active learning frameworks. *Proceedings of the 12th Australasian Data Mining Conference (AusDM 2014)*. Brisbane, Asutralia, 2014 [In Press].

10. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. San Francisco, CA, USA, 2001:282-89.

11. Pradhan S, Elhadad N, South B, et al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. Online Working Notes of CLEF, CLEF 2013, 2013.

12. Lewis DD, Catlett J. Heterogenous Uncertainty Sampling for Supervised Learning. *Proceedings of the 18th International Conference on Machine Learning*. Williamstown, MA, USA, 1994:148-56.

13. Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008:1070-79.

14. Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. Journal of biomedical informatics 2012;**45**(2):265-72.

15. Rosales R, Krishnamurthy P, Rao RB. Semi-supervised active learning for modeling medical concepts from free text. *Proceedings of the Sixth International Conference on Machine Learning and Applications*. Cincinnati, Ohio, USA, 2007:530-36.

16. Zhang H-T, Huang M-L, Zhu X-Y. A unified active learning framework for biomedical relation extraction. J. Comput. Sci. Technol. 2012;**27**(6):1302-13.

17. Boström H, Dalianis H. De-identifying health records by means of active learning. Recall (micro) 2012;**97**(97.55):90-97.

18. Figueroa RL, Zeng-Treitler Q, Ngo LH, et al. Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association 2012;**19**(5):809-16.

19. Sang EFTK, Veenstra J. Representing text chunks. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Bergen, Norway, 1999:173-79.

20. Nguyen AN, Lawley MJ, Hansen DP, et al. A simple pipeline application for identifying and negating SNOMED clinical terminology in free text. *Proceedings of the Health Informatics Conference (HIC)*. Canberra, Australia, 2009:188-93.

21. Culotta A, McCallum A. Reducing labeling effort for structured prediction tasks. *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2005:746–51.

22. McCallum AK. MALLET: A Machine Learning for Language Toolkit. Secondary MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu.

23. Settles B, Craven M, Friedland L. Active learning with real annotation costs. *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, 2008:1-10.

Figure legends

Figure 1. An example of input text and associated concepts from the i2b2/VA 2010 dataset.

Figure 2. An example of BIO tag representation.

Figure 3. A generic pool-based AL algorithm.

Figure 4. The performance of supervised (Sup), Random Sampling (RS), Longest Sequence (LS), and active learning approaches (LC and ID) in InALCE-Tun setting (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.

Figure 5. 10-fold cross validation results across the batches on i2b2/VA 2010 (gray curve) and ShARe/CLEF 2013 (black curve) datasets. The horizontal axis corresponds to the number of batches used for training and the vertical axis reports F1-measure values. Bars indicating the standard deviation across the folds are reported for each batch along the learning curves (a) RS (b) ID (c) LC.

Figure 6. The distribution of full duplicate sequences based on their length in i2b2/VA 2010 dataset.