Semi-Supervised Active Learning for Modeling Medical Concepts from Free Text

Rómer Rosales
IKM CKS
Siemens Medical Solutions
Malvern, PA 19355 USA

Praveen Krishnamurthy
Department of Computer Science
State University of New York
Buffalo, NY 14260 USA

R. Bharat Rao
IKM CKS
Siemens Medical Solutions
Malvern, PA 19355 USA

Abstract

We apply a new active learning formulation to the problem of learning medical concepts from unstructured text. The new formulation is based on maximizing the mutual information that a sample labeling provides about the retrieval/classification model. This methodology is related to and extends the Query-by-Committee approach (QBC) [12] by exploiting unlabeled data in novel ways, beyond their common use only as potential query points. Unlike QBC, this method allows us to employ unlabeled data in addition to labeled data in order to select more appropriate samples for labeling. The samples thus chosen are both informative and also relevant according to a distribution of interest. This flexibility allows us to also tailor the model to arbitrary distributions relevant to the task at hand, in particular to the distribution of the test data. This formulation has implications in scenarios where the training and test distributions are different, or when a general model is adapted to a more specific model. Experiments were conducted to evaluate retrieval performance of natural-language text associated to various concepts of interest in the medical domain. We demonstrate the advantages of our formulation compared with QBC, the state-of-the art active learning approach, and against random sample selection.

1 Introduction

Concept learning can benefit considerably from active user interaction or feedback. For example, it has been observed that Information Retrieval (IR) system interfaces that support interactive collaboration can significantly increase retrieval effectiveness [5]. Likewise, the learning and knowledge discovery process can, in many cases, be made more controlled and efficient by user guidance.

In active learning [8, 9, 12], also called *experimental design* in statistics, unlabeled or unclassified data is available to the algorithm. At each step, the algorithm must appropriately choose an example for a user to label. The final objective is that of learning the appropriate concept using the least number of labels. In knowledge discovery, a system must make efficient use of vast volumes of data to map it into the underlying patterns; one way to address this is by formulating optimized requests for user guidance. The above examples of user feedback and guidance can in general be abstracted by employing the concept of labeling.

In many learning tasks, it is often the case that the labeled data is limited in quantity or expensive to obtain, but the amount of unlabeled data is large or easy to obtain. This applies to many data mining, information retrieval, and general classification and regression problems, but nowadays this is particularly evident in text-based language processing tasks. In the medical domain, text documents are abundant (*e.g.*, in patient records, lab reports, etc) but only a few are or can be labeled with a concept or topic of interest, mainly due to the high cost of labeling, usually requiring specialized knowledge.

In this paper we use text documents as our data source, but the following formulation applies to other information sources. In particular, we apply these ideas to medical informatics by exploring and modeling medical concepts commonly found in electronic medical records.

Active Learning Theory It has been shown that the number of data points needed for learning some functions can be reduced drastically (exponentially) if these points are chosen appropriately. For a class of noiseless, deterministic classification problems, active learning requires $O(\log(1/\epsilon))$ labels to find the classification boundary guaranteeing ϵ error while passive learning requires $O(1/\epsilon)$ examples [4]. While strict error bounds like the above can be analytically obtained only for a limited class of problems, empirical evidence suggests that active learning can be efficient in more practical scenarios (*e.g.*, see [2, 10]).

Related Approaches For analysis purposes, active learning (AL) methods can be divided into a few classes. Active learning by *uncertainty sampling*, such as [7, 2], is the process of selecting the unlabeled data point whose label has highest uncertainty given the current model. An unfortunate effect of this criterion is that noisy or rare data points

tend to be chosen as the model cannot describe them appropriately; these points are often not very useful for building a classifier. A different criterion is provided in [11] where the data point of choice is that which, when labeled, minimizes the estimate of expected (future) error. While this criterion attempts to directly optimize performance, an analytical expression for the expected error rarely can be obtained, and sampling needs to be employed. However, appropriately sampling from a distribution of interest is by itself a difficult task in practice. Moreover, this method requires retraining the model for every point to be considered; this can be computationally prohibitive.

Query-by-Committee (QBC) methods, whose fundamental concept was introduced in [12], and further developed in [4], offers a different perspective. In QBC, the data point that reduces the *size of the version space*[1] is selected. It turns out that this optimal point can be approximated by that for which a set of independently trained models disagrees the most (*i.e.*, regarding its label) (thus the term *committee*). This concept has been embraced by various formulations, including some examples in text processing and information retrieval [10, 1]. In QBC, the model only needs to be evaluated, not retrained, for every competing data point.

Comparison with presented approach Our approach is motivated by QBC, but it extends it in several important ways that allow the use unlabeled data in novel roles (besides their use only as potential query points[2]). QBC can be derived from information theoretic principles. Thus, in the following section, we start by defining a criterion for active learning involving the maximization of Mutual Information [3] relating various random variables defined over data points, labels, and model parameters. We show that when we condition this expression on appropriate random variables, we obtain a range of problems that extends QBC by allowing us to use active learning to more appropriately employ the unlabeled data. This allows us to (1) select points that are representative of the underlying data distribution and that are at the same time *informative* and, as a related effect, (2) aim the descriptive power of the probabilistic model towards arbitrary (desirable) regions of the data space, such as that of known test points.

2 Unlabeled Data and Active Learning

The majority of active learning approaches pay little or no attention to the role of unlabeled data points beyond that of being potential query points. In other words, the score that the active learning algorithm assigns to a potential query or sample point is independent of the other unlabeled points. An exception of the above is the recent transductive experimental design formulation given in [17], which applies transductive learning (and thus uses all unlabeled data points) to active learning for the case of regression problems. This distinction between a passive vs. more influential role of the unlabeled data points is fundamental for the formulation in this paper.

The use of unlabeled data points motivated the concept of semi-supervised learning. We believe that unlabeled data can also play an important role in the context of active learning. We show that by using the information provided by all or part of the unlabeled data, it is possible to improve retrieval/classification performance. The present formulation does not restrict the definition of unlabeled data to just those data points that can be labeled by a user (query points). Specifically, we can select any subset of the unlabeled points, such as the test set (in cases where this is known beforehand), and concentrate the model descriptive power towards these, potentially more relevant, data points.

This is a more general problem since for diverse reasons it may be advantageous to aim the modeling power at particular regions of the space. There has been increasing interest in problems derived from having a test data distribution that is known to differ from the training data distribution [13, 6]. This is also of interest for handling concept drift or non-stationary data distributions.

In active learning, even if the points of interest are known beforehand, it is unclear what criteria should be used that incorporates such points in the formulation (*i.e.,* that incorporates the unlabeled test points in the unlabeled point selection problem); however, see [14] for a method to address linear regression using the expected generalization error. Our formulation shows how the above problem can be approached using general information theoretic principles.

3 Formulation

Let Y be a random variable representing the incidence of a particular concept of interest; Y is defined over the domain $\mathcal{Y}$ (*e.g.,* $\mathcal{Y} = \{\text{true,false}\}$). Let X be a random variable defined over our input data representation space $\mathcal{X}$. For example, the input representation can be a text string itself but usually a more compact representation can be employed. We use a subscript to denote a particular random variable (*e.g.,* X_i) where i belongs to some index set $\mathcal{D}$. For groups of random variables, we use sets as subscripts (*e.g.,* $X_\mathcal{S}$, with $\mathcal{S} \subset \mathcal{D}$).

We define a model of a concept and text strings (or any other input representation) as a probability distribution p[3]. To this end, $p(Y = \mathbf{y}, X = \mathbf{x}, \Theta = \theta)$ denotes the joint

[1] A measure representing the number or volume of parameters that are consistent with the data

[2] Points for which a label can be requested.

[3] We use text strings as an example, since in our experiments we will focus on unstructured text data, but other representations can be used, including combination of text and categorical data.

probability of concept incidence $\mathbf{y}$, input $\mathbf{x}$, and model (parameters) θ. Throughout this paper we will use the notation $p(\mathbf{x})$ to denote $p(X = \mathbf{x})$ for any random variable or set of variables X.

3.1 Active Learning with Unlabeled Data

We consider exploiting our knowledge of available, unlabeled data points in a novel way. As a consequence, the information provided by labeling a data point $\mathbf{x}_i$ will now depend not only on the data point in question, but on the rest of the unlabeled data (and the model parameters as usual).

We represent every data point as a random variable. For labeled data point i, we associate the random variable pair X_i, Y_i taking values denoted $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively. If the point is unlabeled, then only $X_i = \mathbf{x}_i$ would be observed, while Y_i is unobserved (Y_i is a hidden random variable).

Let us define the general problem as that of finding the unlabeled data point whose label, once observed, provides the maximum mutual information about the model θ, conditioned on everything that has been observed (labeled data points with their labels and unlabeled data points). This can be formalized by the following optimization problem:

$$\arg\max_{i \in \mathcal{U}} \quad I(\theta; Y_i | X_{\mathcal{D}} = \mathbf{x}_{\mathcal{D}}, Y_{\mathcal{L}} = \mathbf{y}_{\mathcal{L}}), \quad (1)$$

where $\mathcal{U}$ and $\mathcal{L}$ denote the sets of unlabeled and labeled data points respectively, and $\mathcal{D} = \mathcal{U} \bigcup \mathcal{L}$ denotes all the data points in the dataset. This is equivalent to:

$$\begin{aligned} & \arg\max_{i \in \mathcal{U}} H(Y_i | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) - H(Y_i | \theta, \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \\ = & \arg\max_{i \in \mathcal{U}} - \sum_{\mathbf{y}_i} p(\mathbf{y}_i | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \log p(\mathbf{y}_i | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \\ & + \int_{\theta} \sum_{\mathbf{y}_i} p(\mathbf{y}_i, \theta | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \log p(\mathbf{y}_i | \theta, \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) d\theta \quad (2) \end{aligned}$$

where H denotes the entropy of a random variable[3].

Since in general the integral over θ cannot be solved analytically, we will restrict θ to take values in a finite domain with cardinality K (K can be as large as desired or appropriate to given computational resources). We will index the values of theta by the set $\{1, ..., K\}$. This normally simplifies the computational needs in the above problem to:

$$\begin{aligned} \arg\max_{i \in \mathcal{U}} & - \sum_{k=1}^{K} \sum_{\mathbf{y}_i} p(\mathbf{y}_i, \theta_k | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \log \sum_{j=1}^{K} p(\mathbf{y}_i, \theta_j | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \\ & + \sum_{k=1}^{K} \sum_{\mathbf{y}_i} p(\mathbf{y}_i, \theta_k | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \log p(\mathbf{y}_i | \theta_k, \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}). \quad (3) \end{aligned}$$

This be expressed using the conditional KL divergence[3][4]:

$$i^* = \arg\max_{i \in \mathcal{U}} \mathrm{KL}[p(\mathbf{y}_i | \theta, \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) || \sum_{j=1}^{K} p(\mathbf{y}_i, \theta_j | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}})], \quad (4)$$

where the divergence is computed over the domain of both $\mathbf{y}_i$ and θ. Note that the domain of θ is the set $\{\theta_1, ..., \theta_K\}$, and $\mathbf{x}_{\mathcal{D}}$, $\mathbf{y}_{\mathcal{L}}$ are known, fixed observations (they represent the given data points and known labels).

The parameter values $\{\theta_1, ..., \theta_K\}$ are yet to be specified. Generally speaking, if the probability $p(\mathbf{y}_i, \theta | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}})$ does not have sufficient mass at these values, then the approximation in Eq. 3 would be inaccurate.

Since $p(\mathbf{y}_i, \theta | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) \propto p(\theta | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}}) p(\mathbf{y}_i | \theta, \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}})$, we would like to obtain values of θ that are representative of $p(\theta | \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{L}})$ (*i.e.*, we would like to sample from this posterior). One way to achieve this is by by choosing random training examples from the dataset multiple times (in our case K times)[5]. Each time, this random subset of training points can be used to compute a (Maximum-a-Posteriori or MAP) estimate θ_k with $k = \{1, ..., K\}$ (see algorithm box).

Algorithm for Semi-supervised Active Learning
Inputs: $\mathbf{x}_{\mathcal{U}}, \mathbf{x}_{\mathcal{L}}, \mathbf{y}_{\mathcal{L}}$ (data), K.
Output: K models trained with actively labeled data.

1. Label a few randomly chosen data points
2. For each model k compute MAP estimate for θ_k by randomly assigning labeled data points to each model's training set
3. Repeat until a stopping criteria is met
   (a) For each unlabeled data point i (whose label could potentially be requested)
   - Compute KL divergence score using Eq. 4

   (b) Select next point to label/query by finding the point i with maximum KL divergence
   (User feedback: Point is labeled by user)
   (c) For each model k compute MAP estimate for θ_k using a random selection of labeled points

3.2 Relation to Query-by-Committee (QBC)

This is related to the Query-by-Committee (QBC) formulation [12]. In a practical implementation of QBC, the procedure is to choose the point for which a set of models (committee) disagrees the most with respect to its label, also

[4]For discrete random variables x and y, the conditional KL divergence is defined $\mathrm{KL}(p(y|x)||q(y|x)) = \sum_{x,y} p(y,x) \log[p(y|x)/q(y|x)]$

[5]We do not need to obtain samples for $\mathbf{y}_i$ since we are summing over its full domain

in terms of the KL divergence. Formally:

$$i^*_{QBC} = \arg\max_{i \in \mathcal{U}} \sum_{k=1}^{K} \mathrm{KL}[p(\mathbf{y}_i|\theta_k)||\frac{1}{K}\sum_{j=1}^{K} p(\mathbf{y}_i|\theta_j)] \quad (5)$$

We can show that the above objective function may be obtained by maximizing the mutual information, but without conditioning on observations. Also, note that Eq. 4 is written as a conditional KL divergence [3] defined over the label and model variables and not a sum of KL divergences defined over the label random variable alone, as in QBC. Likewise, in the new formulation the sum in the second term marginalizes over the model space. As a result every model (from the committee) is taken into account differently.

We will see that this extension: (1) opens the door to semi-supervised forms of active learning, where unlabeled data plays a more influential role in the selection of points to label and (2) allows to explicitly direct the model descriptive power through active learning.

4 Semi-supervised Active Learning

The above formulation offers a more general framework for active learning problems, in particular in cases where unlabeled data is available beforehand.

In order to develop the above formulation for a definite family of models, we now consider the graphical model in Fig. 1 representing the following factorization:

$$p(\mathbf{x}_\mathcal{D}, \mathbf{y}_\mathcal{D}, \theta) = \prod_{l \in \mathcal{L}} p(\mathbf{x}_l|\mathbf{y}_l, \theta) p(\mathbf{y}_l|\theta) \prod_{u \in \mathcal{U}} p(\mathbf{x}_u|\mathbf{y}_u, \theta) p(\mathbf{y}_u|\theta) p(\theta).$$

The important property of this model, relevant to our analysis, is the conditional independence of the random variable pairs (X_i, Y_i) and (X_j, Y_j) given the model parameters θ for any (i, j), labeled or unlabeled. Algorithms for (supervised) training are standard, thus we do not cover this here. However, for efficiently computing the KL divergence it is useful to note that $p(\mathbf{y}_i|\theta_j, \mathbf{x}_\mathcal{D}, \mathbf{y}_\mathcal{L}) = p(\mathbf{y}_i|\theta_j, \mathbf{x}_i)$ and $p(\theta_k|\mathbf{x}_\mathcal{D}, \mathbf{y}_\mathcal{L}) = p(\theta_j|\mathbf{x}_\mathcal{U})$. We defer the derivations due to space limitations[6].

4.1 Aiming Model Descriptive Power

While throughout the presented formulation $\mathbf{x}_\mathcal{U}$ has represented the set of unlabeled data points, normally referred to as those that could be potentially labeled by the user, we can also assign other meanings to this set. Note that the selection criterion implied by our active learning formulation attempts to query points that are informative (in the usual QBC sense) but also representative of the underlying data distribution. The role of this unlabeled set $\mathbf{x}_\mathcal{U}$ is to provide an approximation of the underlying data distribution.

[6]This is easy to show by using the conditional independence assumptions implied by the Bayes network

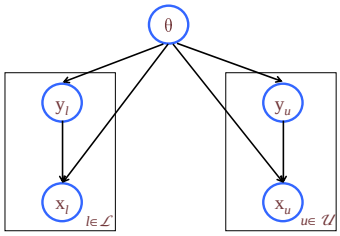

Figure 1. General Bayesian network structure employed for the analysis in Sec. 4, where (X_i, Y_i) is independent of (X_j, Y_j) given the model θ. We use the most basic representation for the conditional distribution $p(\mathbf{x}_i|\mathbf{y}_i, \theta)$ (for any i), but this distribution can be arbitrary, and include other variables, as long as the independence is maintained.

The above formulation remains the same if for some reason we desire to assign to this set only part of the unlabeled data or any other set of data points. One particular situation of great interest is the case when there is knowledge of the types of data points that the model could encounter during testing (*e.g.*, at classification or retrieval). This allows us to tailor active learning to particular test data sets. The goal is to obtain models that achieve high performance in the test dataset by leveraging knowledge of the test data distribution. An application of this concept will be seen in the experiments.

We remark that a natural use of this property is for approaching problems where train and test distributions are different [13]. Another type of problems that can benefit from this approach is model refinement. This is the case where a generic model was estimated using generic data, but once more specific data become available, active learning can be geared towards better modeling such data.

5 Experimental Evaluation

Experiments were performed using actual electronic medical records (EMR) from patients at a medium-sized heart hospital[7]. Our experiments are designed to evaluate the proposed methodology in our ultimate objective of correct retrieval and classification of patients/documents with certain medical conditions or state from unstructured text (*e.g.*, transcribed doctor notes, lab reports, etc.).

5.1 Representation and Datasets

We designed our experiments to work at the sentence level; thus in reference to our notation in Sec. 3, X_i is a random vector representing a sentence-based observation.

[7]Name undisclosed due to privacy agreements.

With the help of expert personnel, we concentrated on gathering information about six medically relevant concepts related to heart disease. These concepts were chosen primarily due to their prevalence in medical records involving heart related diagnosis and treatment. Basic properties of the datasets are shown in Table 1.

These sentences were obtained from a set of $\sim$2 million sentences by searching for one or two keywords obtained simply from the concept name. A random subset of the matching sentences were labeled by an expert and saved. As expected, keywords were only useful at a first level retrieval; not surprisingly a mixture of completely irrelevant, affirmative, or negative sentences were obtained; *e.g.,* not all patients with documents containing the keyword *smok** are actual smokers or even have a history of smoking, as seen in the table.

In our task, the labels T=True and F=False were chosen to indicate the following: (1) T $\rightarrow$ the concept is *present and affirmative* in the sentence (2)F $\rightarrow$ the concept is *absent* OR the concept is *present but negated* in the sentence. In general one can further divide the F label into F (*present but negated*) and N/A (*absent*).

5.2 Experimental Settings

We set the number of committee members $K = 15$ throughout all our experiments. The data was used as follows: for each dataset, we first divided it into two subsets, one held out for testing only (20%) and one used for training (80%). From the training subset, a portion was assigned for initial training (8%)(each model was initially trained without active learning with a different random set of points from this portion), and the remaining examples (92%) were assigned for active learning; this last set is formed by the points that the active learner can query (request the user to label). Finally, not all active learning sentences are used, but only a fixed amount (75% of these). These settings were the same for all datasets. A total of 10 runs were performed where the above subsets were always randomized. The method does not require any tunable parameter other than K. The performance reported is the average across all K models. In case an optimal model need to be chosen among the above K, a cross-valuation set can be employed.

In order to test our approach without any further special domain knowledge tuning, which can influence the results considerably, we used a naive Bayes network as our probabilistic model. In practice any probabilistic model can be used, as long as it satisfies our iid assumption in Sec. 4. The network had a fixed number of input features (words in our case), set to 20. The words were chosen in order to maximize the mutual information between the words appearance random variable and the label. This calculation was part of model training. No specialized medical or general linguistic databases (*e.g.,* [15, 16]) were used. This adaptation goes beyond the scope of this work[8].

Table 1. Test Datasets

Name	# Labeled	(%T/%F)
1. Tricuspid Valve Replacement	255	(45%/55%)
2. Mitral Regurgitation	321	(26%/74%)
3. Assisted Living	123	(60%/40%)
4. Congestive Heart Failure	264	(81%/19%)
5. Smoking Currently	383	(49%/51%)
6. Smoking History	383	(73%/27%)

5.3 Results

In order to carefully test our formulation, we compared two embodiments of our method against random sentence selection for labeling (*i.e.*, label any randomly chosen sentence) and against QBC sentence selection (the standard information theoretic active learning approach).

While the two versions attempt to solve Eq. 1 (using the approximation in Eq. 4), the set $\mathbf{x}_{\mathcal{U}}$ is different for the two versions. In version 1 $\mathbf{x}_{\mathcal{U}}$ are data points in the active set. In version 2, $\mathbf{x}_{\mathcal{U}}$ are data points in the test set. In any case, the points available for querying/labeling are only those in the active set.

We measured two basic performance quantities (1) relative remaining error on the active set (2) relative remaining error in the test set. Performance error (E) is defined as the number of incorrect retrievals (equivalently labeling a sentences as F when its label is T and *vice versa*). The relative remaining error is the proportion of error still remaining after running the algorithm, compared to the initial error: $RE = E_f/E_0$, where E_f is the performance error after training and E_0 is the performance error after step (2) of the algorithm in Sec. 3.1. We use the RE measure to eliminate any performance advantages after step (2) that may have resulted from randomly sampling the dataset. The error in the active set is calculated on the remaining unlabeled points after each step. Randomized labeling is shown for completeness, but was clearly less effective overall. Fig. 2 shows the combined results indicating performance in test and active sets, for all algorithms. Mean and standard deviations reported are with respect to 10 (randomized) runs (see Sec. 5.2) and K models. Versions 1 and 2 are referred to as [Cond on active set] and [Cond on test set] respectively.

Regarding the performance on the test set, we observed that the two versions of our algorithm outperform QBC in average. This result was observed for all datasets. Our method conditioned on the test set always performed better

[8] Additionally, it is not clear how to appropriately incorporate these specialized domain knowledge sources into an information theoretic active learning formulation

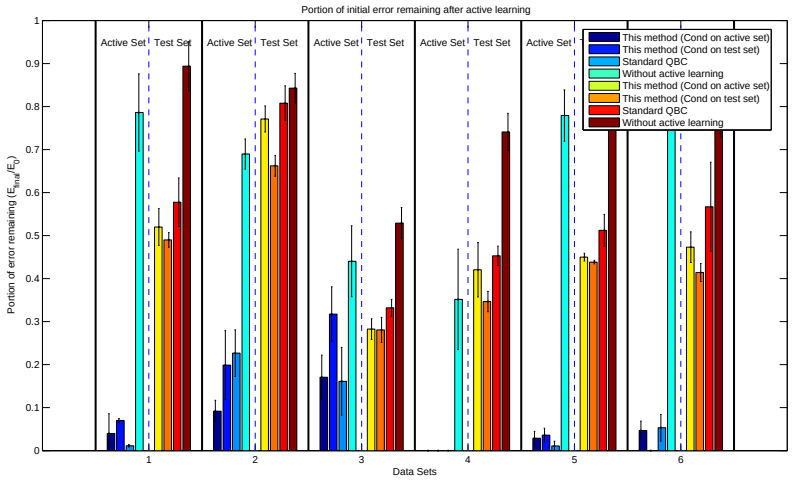

Figure 2. Relative performance comparison between (1) the proposed method conditioned on the active set (examples that can be requested for labeling) (2) proposed method conditioned on the test set (held out, not available for labeling), (3) QBC, and (4) random selection. Each algorithm is tested on all datasets. Relative performance is computed on active (left) and test (right) sets.

in average than our method conditioned on the active set. This clearly demonstrates that it is possible for the active learning formulation to improve performance in an arbitrary set (in particular a test set). The difference in performance between our method conditioned on the test set and QBC was always statistically significant.

Regarding the performance on the active set, we observed that in general, our method conditioned on the active set performs better than our method conditioned on the test set (except for dataset 6, where surprisingly the error was reduced to almost zero by the second algorithm). This result is consistent with our formulation and indicates (along with the previous result) that the conditioning on different sets have the desired behavior. We could not observe a clear difference between QBC and our method conditioned on the active set, but note that in most cases the error decreases to zero, making comparison less relevant.

Fig. 3 shows the percentage incorrectly labeled sentences as a function of the number of training examples provided. The mean curves are shown, computed across 10 runs. Note that there are four curves: two for QBC and two for our method (conditioned on the test set). One curve represents the error measured on the test set and the other on the active set. We can observe that the error evolution on the test set tends to level after a number of active learning steps, but we expect the error to increase after more steps as a common consequence of over-fitting. In all cases the average error in the test set is lower for our formulation compared with QBCs error in the test set. This is expected since our method aims its descriptive power to the test set while the standard QBC does not. Regarding the active set, our model's performance is comparable with QBC's since we are not aiming the descriptive power to this set.

The rate at which error is reduced is comparable for both methods (in all scenarios). This is not a drawback in our formulation since the bound on the number of labels required by QBC is comparable to the best known bounds. The error on the active set tends to decrease to zero for both methods. We believe this is because only the *easy* examples are left in the active set after a number of iterations.

6 Conclusions

We have approached the problem of medical concept learning from text using a novel formulation for active learning. The formulation is designed to incorporate unlabeled data points in a new manner. It is thus, unique at extending QBC by its handling of unlabeled data. We have shown clear performance advantages over the widely used QBC when evaluating the results on a held out test set. More importantly, our method can be employed to aim the model descriptive power to arbitrary areas of the input space (*e.g.*, arbitrary set of documents). We have pointed out connections to QBC, addressing when they are equivalent. The resulting formulation has a natural interpretation: it proposes to choose data points that are informative but also representative of the distribution of interest.

There exist practical problems derived from having a test distribution that differs from the available training distribution; a topic that is receiving increasing attention in artificial

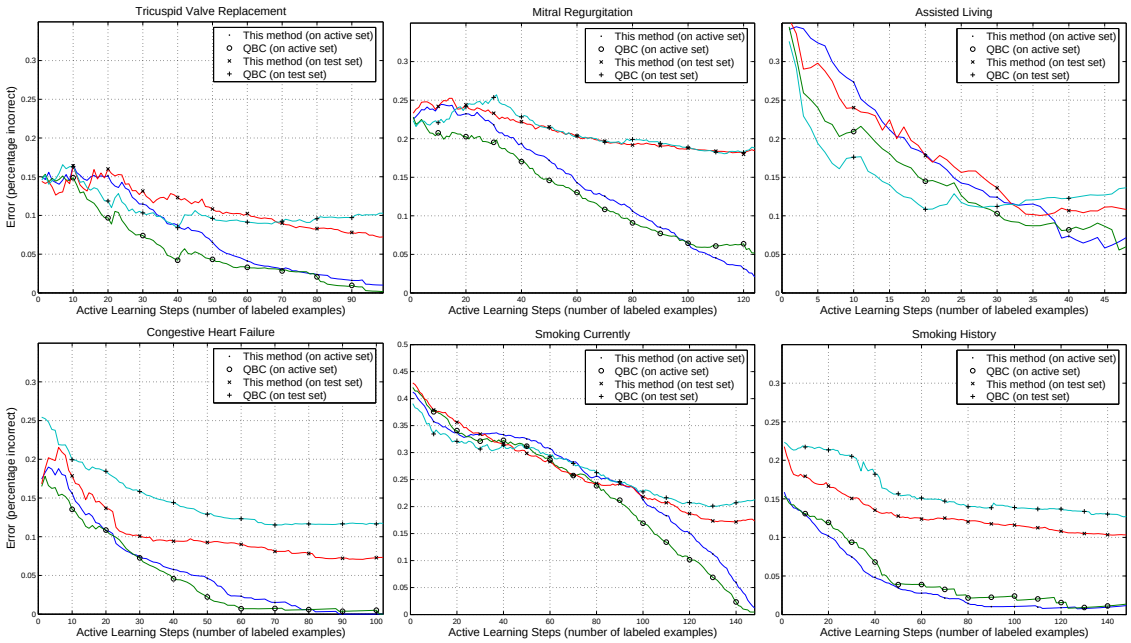

Figure 3. Percentage incorrect labels and number of actively labeled samples for our formulation vs. QBC. Performance for test and active sets is shown. Curve was sampled at every active learning step (iteration); however to avoid cluttering the curve corresponding marker $(\cdot, \circ, \times, +)$ is shown every 10 active learning steps.

intelligence. In medical informatics, there are cases where a training corpus exist for some particular domain, but the actual domain of the task varies. This has been our experience when data from one medical institution or health care provider has been labeled (at a high cost), but we wish to apply our learned model to data from a different medical institution. In a more general sense, we believe it may be beneficial to use a training corpus to learn generic models, which can then be fine-tuned through our active learning approach by conditioning on data from the new, more specific task. This is also of interest for handling concept drift or non-stationary data distributions. We see our formulation as a useful tool that can be exploited toward approaching these interesting problems.

References

[1] B. Anderson and A. Moore. Active learning for hidden markov models: Objective functions and algorithms. In *International Conference on Machine Learning*, 2005.

[2] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *J. of Artificial Intelligence Research*, 4:129–145, 1996.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, 1991.

[4] Y. Freund, S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 2-3:133–168, 1997.

[5] J. Koenemann and N. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. *CHI*, pages 205–212, 1996.

[6] N. Lawrence, A. Schwaighofer, J. Quinonero-Candela, and M. Sugiyama. Workshop on learning when test and training inputs have different distributions. NIPS, December 2006.

[7] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12, 1994.

[8] D. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Stat*, 27:986–1005, 1956.

[9] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4:590–604, 1992.

[10] A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Int. Conf. Machine Learning*, 1998.

[11] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Int. Conf. Machine Learning*, pages 444–448, 2001.

[12] S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *W. Comp. Learning Theory*, pages 287–94, 1992.

[13] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. of Statistical Planning and Inference*, pages 227–244:90(2), 2000.

[14] M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *J. Mach. Learning Research*, pages 141–166:7, 2006.

[15] Unified medical lang. system. http://umlsinfo.nlm.nih.gov/.

[16] WordNet. http://wordnet.princeton.edu/.

[17] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Int. Conf. Machine Learning*, 2006.