External Knowledge and Query Strategies in Active Learning: a Study in Clinical Information Extraction

Mahnoosh Kholghi[1,2], Laurianne Sitbon[1], Guido Zuccon[1], Anthony Nguyen[2]
[1]Queensland University of Technology (QUT), Brisbane 4000, Queensland, Australia
[2]The Australian e-Health Research Centre, CSIRO, Brisbane 4029, Queensland, Australia
{m1.kholghi, laurianne.sitbon, g.zuccon}@qut.edu.au, anthony.nguyen@csiro.au

ABSTRACT
This paper presents a new active learning query strategy for information extraction, called Domain Knowledge Informativeness (DKI). Active learning is often used to reduce the amount of annotation effort required to obtain training data for machine learning algorithms. A key component of an active learning approach is the query strategy, which is used to iteratively select samples for annotation. Knowledge resources have been used in information extraction as a means to derive additional features for *sample representation*. DKI is, however, the first query strategy that exploits such resources to inform *sample selection*. To evaluate the merits of DKI, in particular with respect to the reduction in annotation effort that the new query strategy allows to achieve, we conduct a comprehensive empirical comparison of active learning query strategies for information extraction within the clinical domain. The clinical domain was chosen for this work because of the availability of extensive structured knowledge resources which have often been exploited for feature generation. In addition, the clinical domain offers a compelling use case for active learning because of the necessary high costs and hurdles associated with obtaining annotations in this domain. Our experimental findings demonstrated that (1) amongst existing query strategies, the ones based on the classification model's confidence are a better choice for clinical data as they perform equally well with a much lighter computational load, and (2) significant reductions in annotation effort are achievable by exploiting knowledge resources within active learning query strategies, with up to 14% less tokens and concepts to manually annotate than with state-of-the-art query strategies.

Categories and Subject Descriptors I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

General Terms Algorithms, Performance, Experimentation.

Keywords Domain Knowledge; Active Learning; Concept Extraction; Clinical Free Text; Conditional Random Fields.

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.

ACM 978-1-4503-3794-6/15/10…$15.00
http://dx.doi.org/10.1145/2806416.2806550

1. INTRODUCTION
Supervised machine learning (ML) algorithms can train powerful statistical models to effectively capture useful structured information from clinical free-text resources [19]. However, these algorithms require a large amount of annotated training data. Preparing high quality annotated data is a major obstacle to effective data analysis in many domains; this is even further exacerbated in the clinical domain as it requires significant efforts from highly trained, and costly, clinical professionals.

Active learning (AL) has been suggested as an effective solution for reducing the burden of manual annotations in automatic information extraction tasks [26]. In standard supervised learning, training samples are annotated at once. Active learning differs from this because the annotation is performed as an iterative process in which *informative* samples (i.e., those that contain useful information for the model) are selected from a pool of unlabeled data and provided to experts for annotation. The selection of informative samples takes place according to a query strategy, which thus represents a key component within AL approaches. In each iteration, after samples are selected and annotated, they are added to the labeled set of data, on which a refined supervised model is built [26]. It is possible to either fully retrain the model at each step, or incrementally update the model learnt in the preceding loop (see Figure 1).

AL approaches vary in the query strategies that are adopted, as well as in the supervised machine learning algorithm that is

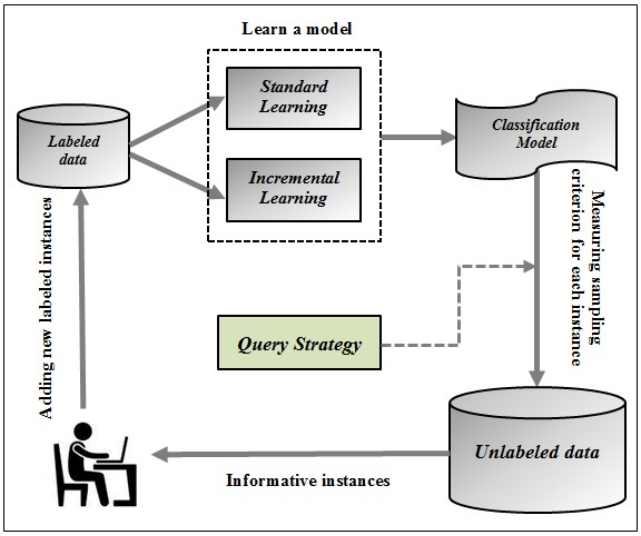


Figure 1. The active learning process.

employed throughout the iterations; as a result, AL has shown varying performance across different datasets [6].

Despite its practical benefits, active learning is not fully explored in clinical information extraction [19], in particular for the task of clinical concept extraction, which represents the focus of this paper. This task involves capturing natural language sequences (words or multi-words expressions) belonging to pre-determined semantic categories from unstructured free texts. These terms express meaningful concepts within a given domain (e.g. clinical problems, tests, and treatments) [8, 9]. Extracting clinical concepts is thus an important primary step in identifying meaningful information from clinical free texts.

The clinical domain is rich in terms of information resources. Clinical knowledge resources (e.g., the UMLS[1] [16] and the SNOMED CT[2] [4]) and tools (e.g., MetaMap [1], MedLEE [7], MedEx [35], cTAKES [24], and Medtex [18]) have been widely leveraged in combination with dictionary-, rule- and machine learning-based approaches in order to improve clinical concept extraction. The strong need for effective information extraction methods in the clinical domain has encouraged the development of shared datasets such as the i2b2 challenges [32-34] and the ShARe/CLEF eHealth Evaluation Lab [10, 31], which in turn have sparked the development of novel, more effective clinical information extraction methods. Specifically, a wide attention has been given to the engineering of powerful features obtained from domain knowledge resources to strengthen the supervised machine learning models and increase their effectiveness [21, 34].

While these resources have been used for data representation (i.e., features) both in supervised and active learning approaches [11], there is no study investigating the contributions these resources could have to improve current state-of-the-art query strategies within active learning methods for clinical information extraction.

In this paper, we address this gap and propose a novel active learning query strategy, called Domain Knowledge Informativeness (DKI), which leverages domain knowledge together with informativeness measures to select those samples that most strengthen the model obtained at the previous active learning iteration. We investigate the comparative performance of DKI and a wide range of state-of-the-art query strategies in terms of annotation effort savings in the clinical domain. The clinical domain is chosen for this work because of the already discussed availability of extensive structured knowledge resources. It also offers a compelling use case for active learning because of its intrinsic high costs and hurdles associated with obtaining annotations. To the best of our knowledge, DKI is the first query strategy which incorporates domain knowledge when querying samples within the active learning framework. The research questions we seek to tease out in this work are:

1. Which non-knowledge based query strategy is better suited to clinical data?
2. Can domain knowledge support a more effective query strategy that further reduces annotation efforts compared to state-of-the art AL approaches?

We find that, when comparing the current state-of-the-art query strategies, as well as a few novel strategies derived from existing ones for active learning, the confidence about the label of a sample as estimated by the classification model is a key factor for discovering informative samples in the clinical domain.

Our findings also suggest that domain knowledge can play an important role for enhancing active learning's performance by further reducing annotation effort. We also bring forward insights about how other machine learning approaches, such as semi-supervised learning, can be used to augment active learning.

The rest of this paper is organized as follows: Section 2 presents the problem definition. Section 3 introduces the query strategies. Section 4 describes our experimental and evaluation settings. Results are reported in Section 5. Section 6 briefly reviews related work and Section 7 concludes the paper by outlining directions of future investigation.

2. PROBLEM DEFINITION

Concept extraction (often referred to as entity extraction) can be modelled as a sequence labeling task. In this task, a label sequence $\vec{y} = (y_1, \dots, y_n)$ needs to be assigned to each observed sequence $\vec{x} = (x_1, \dots, x_n)$ in the dataset. Supervised machine learning models are applied to this task by casting the entity recognition problem to that of estimating the posterior probability of $\vec{y}$ given $\vec{x}$ under the model parameters θ. The output sequence with the highest posterior probability is the one that is chosen by the supervised models to annotate the input sequence.

Active learning approaches use supervised machine learning algorithms in an iterative process, where samples[3] from the dataset are successively selected to be annotated by an expert based on their "*informativeness*". Generally, samples are informative if they contain more useful information for the model compared to the rest of the samples in the unlabeled set. The intuition is that identifying and adding informative samples to the labeled set would lead to training a model that would achieve the highest effectiveness. One of the main scenarios in active learning is the pool-based approach (see Figure 2).

Under this paradigm, the active learning system has access to a pool of unlabeled data and, based on a query strategy, the system selects a batch of samples within successive interaction loops to add to the training set. This annotated data is then added to the training set and used to retrain the model [26]. A core issue when

Input:
$\mathcal{L}$: set of labelled samples
$\mathcal{U}$: set of unlabelled samples
$\emptyset$: classifier model
$\varphi(\vec{u}, \emptyset)$: query strategy where $\vec{u} \in \mathcal{U}$
B: number of samples to be selected in each iteration (batch size)

Procedure:
1- Randomly select an initial labeled set $\mathcal{L}$
2- Train a model $\emptyset$ on $\mathcal{L}$;
3- For all samples in the unlabeled set $\vec{u} \in \mathcal{U}$, calculate $\varphi(\vec{u}, \emptyset)$
4- Select a batch of samples (B) where $arg\,max_{\vec{u} \in \mathcal{U}}\ \varphi(\vec{u}, \emptyset)$ and ask the expert to label them;
5- Add samples from step 4 to labeled set $\mathcal{L}$ and remove them from the unlabeled set $\mathcal{U}$.
6- Repeat step 2 to 5.

Figure 2. Pool-based active learning algorithm.

[1] Unified Medical Language System

[2] Systematized Nomenclature of Medicine Clinical Terms

[3] In our experiments, a sample corresponds to a sequence of tokens or a sentence.

designing an active learning framework is: what query strategy should be used to estimate the informativeness of the samples that will be used to retrain or update the classification model? We discuss query strategies for active learning next.

3. Query Strategies

Active learning query strategies for sequence labeling tasks can be categorized in 3 groups of approaches: informativeness based, informativeness-similarity based, and model-independent. Within these categories we included a number of variations to commonly used query strategies (IDiv, IDD, MRD, and ALC).

In addition, we argue that a fourth group is formed by external knowledge-informed approaches, like DKI.

3.1 Informativeness Based Approaches

This group of approaches considers the uncertainty of a model about the label of a sample as a measure of informativeness. These approaches query samples where the learnt model is most uncertain about their label.

3.1.1 Least Confidence (LC)

Least confidence is a common query strategy for measuring informativeness [5]. This query strategy considers the confidence of a model $\emptyset$ with parameters θ about the label $\vec{y}$ of a sample $\vec{x}$. This confidence is estimated based on the posterior probability:

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

The Viterbi algorithm is used to compute the most likely predicted label sequence $\vec{y}^*$.

3.1.2 Margin

Another uncertainty based strategy to measure the informativeness of a sample $\vec{x}$ is to consider the margin between the most likely ($\vec{y}_1^*$) and the second most likely ($\vec{y}_2^*$) label sequences [25]. To calculate the margin, the posterior probability of the two most likely label sequences is subtracted:

$$\varphi_{Margin}(\vec{x}, \emptyset) = -(P_\theta(\vec{y}_1^*|\vec{x}) - P_\theta(\vec{y}_2^*|\vec{x})) \quad (2)$$

The smaller the margin, the more difficult it is for the model to predict the label of a sample. Hence, according to this query strategy, that sample is an informative sample for the model. The first negative sign in Equation (2) ensures that φ_{Margin} acts as a maximizer to be used within the AL algorithm (Figure 2).

3.1.3 Sequence Entropy (SE)

Another way to estimate the informativeness is entropy [30]. Entropy is a measure of uncertainty that indicates the amount of information of a sequence. Sequence entropy [27] is a specialization of entropy for sequence labeling; this is used to find informative samples:

$$\varphi_{SE}(\vec{x}, \emptyset) = -\sum_{\acute{y}} P_\theta(\acute{y}|\vec{x}) log P_\theta(\acute{y}|\vec{x}) \quad (3)$$

where $\acute{y}$ includes all possible label sequences for an sample $\vec{x}$. The higher the entropy, the more informative the sample is.

3.1.4 Augmented Least Confidence

As described in Section 3.1.1, the least confidence (LC) approach uses the model's confidence about a sample's label to find informative samples, i.e., sample selection is based on those that have the lowest probability for the sample's label. To do so, the posterior probability of the model for all samples in the unlabeled set is computed and then samples are ranked accordingly.

The LC approach selects unlabeled samples that are characterized by low classification confidence. However, there are other samples that are characterized by high posterior probabilities, meaning that the model is confident about the samples' label. These samples can be automatically labeled by the model and added to the labeled set in each iteration, so as to provide further learning examples to the classifier. This approach is often referred to as semi-supervised learning [37]. To investigate whether this intuition can positively affect the active learning process, we modified step 4 in the active learning algorithm of Figure 2 as follows:

- Select a batch of samples where $arg\,max_{\vec{u}\in\mathcal{U}}\ \varphi_{LC}(\vec{u}, \emptyset)$ (Equation (1)) and ask the expert to label them. Then select those samples $\vec{u} \in \mathcal{U}$ where $P_\theta(\vec{y}|\vec{u}) > \tau$ (τ is a high probability used as decision threshold) and label them automatically using the current classification model.

We call this approach Augmented Least Confidence (ALC).

3.2 Informativeness-Similarity Based Approaches

The intuition behind informativeness-similarity based approaches is to consider similarity measures for the selection of both representative and diverse samples, in addition to informative ones, aiming to achieve a better coverage of the dataset characteristics.

3.2.1 Information Density (IDen)

Information Density [27] considers the representativeness of samples, along with their informativeness, to prevent outliers to be selected by the active learning process. IDen is computed according to:

$$\varphi_{IDen}(\vec{x}, \emptyset) = \varphi_{informative}(\vec{x}, \emptyset) \times \mathcal{R}_{representative}(\vec{x}) \quad (4)$$

where $\mathcal{R}_{\text{representative}}(\vec{x})$ corresponds to the representativeness of samples:

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

The average similarity between sample $\vec{x}$ and all other samples in the set of unlabeled samples ($\vec{x}^{(U)}$) indicates the representativeness of sample $\vec{x}$: the higher the similarity, the more representative the sample. Similarity is measured according to the cosine distance:

$$sim_{cos}(\vec{x}, \vec{x}^{(U)}) = \frac{\vec{\mathbf{x}}.\vec{\mathbf{x}}^{(U)}}{\|\vec{\mathbf{x}}\|.\|\vec{\mathbf{x}}^{(U)}\|} \quad (6)$$

where $\vec{\mathbf{x}}$ refers to the feature vector of sample $\vec{x}$. To measure the informativeness of samples ($\varphi_{informative}(\vec{x}, \emptyset)$), we use least confidence (Equation (1)). Therefore, according to this query strategy, those samples characterized by the least confidence and the highest similarity are those that are useful to the model.

3.2.2 Information Diversity (IDiv)

Differently from IDen, information diversity aims to take into account the diversity of the data within the process of querying samples; IDiv is formalized as follows:

$$\varphi_{IDiv}(\vec{x}, \emptyset) = \varphi_{informative}(\vec{x}, \emptyset) \times D_{diversity}(\vec{x}) \quad (7)$$

$D_{diversity}(\vec{x})$ is calculated based on how dissimilar a sample $\vec{x}$ is compared to already selected samples within the labeled set $\vec{x}^{(L)}$:

$$D_{diversity}(\vec{x}) = 1 - \left(\frac{1}{L}\sum_{l=1}^{L} sim\left(\vec{x}, \vec{x}^{(L)}\right)\right) \quad (8)$$

Similarity is measured according to Equation (6). LC (Equation (1)) is also used as a measure of informativeness.

3.2.3 Information Density and Diversity (IDD)

One way to find both representative and diverse informative samples is to combine IDen and IDiv approaches as an IDD approach:

$$\varphi_{IDD}(\vec{x}) = \varphi_{informative}(\vec{x}) \times \mathcal{R}_{representative}(\vec{x}) \times D_{diversity}(\vec{x}) \quad (9)$$

3.3 A Model-Independent Approach

Both informativeness and informativeness-similarity based approaches are dependent on the model. In a real world active learning scenario, an expert should wait until a model is learnt on the current labeled set and then the next batch of samples are selected using one of the above-mentioned query strategies for labeling. This could introduce large time delays between expert annotations cycles due to the time required to train models. One way to prevent such a problem is to propose approaches that are independent from the model output.

3.3.1 Maximum Representativeness-Diversity (MRD)

Maximum representativeness-diversity is an approach which only relies on the similarity between a sample $\vec{x}$ and all other samples in the labeled ($\vec{x}^{(L)}$) and unlabeled ($\vec{x}^{(U)}$) sets:

$$\varphi_{MRD}(\vec{x}) = \mathcal{R}_{representative}(\vec{x}) \times D_{diversity}(\vec{x}) \quad (10)$$

$\mathcal{R}_{representative}(\vec{x})$ and $D_{diversity}(\vec{x})$ are calculated according to Equation (5) and (8), respectively. The most representative and diverse samples are labeled in the current batch and then added to the training set.

3.4 External Knowledge-Informed Approaches

External resources and ontologies are useful tools for extracting domain-specific features to represent samples and they have been shown to often enhance supervised machine learning models. However, there are no active learning query strategies that use external knowledge resources to drive the process of sample selection. Here, we tackle this gap and propose a novel query strategy called Domain Knowledge Informativeness.

3.4.1 Domain Knowledge Informativeness (DKI)

DKI is a query strategy to "inform" the model about unlabeled samples using external knowledge. We combine an informativeness based approach with domain knowledge to select samples for active learning:

$$\varphi_{DKI_t}(\vec{x}) = \frac{1}{|\vec{x}|} \times \sum_{\vec{x}} \left(K(x_i) \times \varphi_{informative}(x_i, \emptyset)\right) \quad (11)$$

Where $|\vec{x}|$ is the length of sequence $\vec{x}$, $K(x_i)$ is a function of the domain knowledge contained in the sequence (i.e., importance), and $\varphi_{informative}$ is the informativeness measure. φ_{DKI_t} measures the average importance of the tokens contained in a sequence based on their informativeness for the model as well as the importance of the domain knowledge they carry. A useful domain characteristic that can be extracted from domain knowledge resources is the semantic type that each token belongs to. This information has often been exploited to generate semantic features for sample representation in supervised models. Specifically, here we consider the distribution of semantic types as they appear in the target concept spans in the training set. We use this information to compute the Semantic Value (SV) of a semantic type as follows:

$$Semantic\ Value\ (semantic\ type) = \frac{\#\ semantic\ type\ appears\ in\ annotated\ target\ concept\ span}{\#\ semantic\ type\ appears\ in\ the\ whole\ training\ set} \quad (12)$$

Semantic Value is calculated based on the semantic types assigned to each token individually, without considering neighboring tokens. However, it is possible for a token to be a part of a longer concept span. For example, "*acute*" could be an adjective at the token level, but also part of the longer concept "*acute headache*" which is a disease, and should therefore be considered under this type instead. Since we model the concept extraction problem as a sequence labeling task, instead of considering the semantic types of each token separately, it is more appropriate to calculate the distribution of the semantic type assigned to the longest span that each token belongs to.

In this paper we consider a specific instantiation of DKI, where the domain knowledge contributes to the query strategy by means of the Longest Span Semantic Value (LSSV). Here, LSSV is used to find sequences with the maximum number of possible target concepts:

$$LSSV(semantic\ type) = \frac{\#\ semantic\ type\ appears\ in\ the\ longest\ span\ of\ annotated\ target\ concepts}{\#semantic\ type\ appears\ in\ the\ whole\ training\ set} \quad (13)$$

We use Equation (13) to calculate $K(x_i)$ in φ_{DKI_t} (Equation (11)). $K(x_i)$ is calculated for each token separately ($K(x_i)$, $\vec{x} = \{x_i | i = 0, \dots, |\vec{x}|\}$); informativeness is also measured per token:

$$\varphi_{info}(x_i, \emptyset) = 1 - P_\theta\left(y_j = y_j^* | x_i\right) \quad (14)$$

Where $P_\theta\left(y_j = y_j^* | x_i\right)$ is the marginal probability, y_j^* specifies the label at the corresponding position of the most likely label sequence $\vec{y}^*$.

Note that φ_{DKI_t} is normalized by the length of the sequence $\vec{x}$ ($1/|\vec{x}|$ in Equation (11) is the normalization factor). To encourage the selection of longer sequences as they usually contain more target concepts [27], we also propose a denormalized instantiation of DKI, termed the total token-based Domain Knowledge Informativeness (DKI_{tt}):

$$\varphi_{DKI_{tt}}(\vec{x}) = \sum_{\vec{x}} \left(K(x_i) \times \varphi_{informative}(x_i, \emptyset)\right) \quad (15)$$

4. EXPERIMENTAL SETUP

In this section we introduce the datasets and their preparation steps. We then elaborate on our supervised approach and active learning settings. Finally, our evaluation methodology is explained.

4.1 Dataset Description

We use the annotated training and testing sets for the concept extraction task in the i2b2/VA 2010 NLP challenge [34] and ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) [20].

The i2b2/VA 2010 task requires the extraction of clinical problems, tests and treatments from clinical reports. These reports are a combination of discharge summaries and progress notes. They are organized as a collection of phrases and sentences.

The ShARe/CLEF 2013 eHealth Evaluation Lab (task 1) requires to extract and identify mentions of disorders from clinical free-text notes. The dataset consists of discharge summaries, electrocardiogram, echocardiogram, and radiology reports.

Table 1 reports the number of documents in the training and testing sets of the two datasets along with the number of sequences obtained after pre-processing.

Table 1. Document (#doc) and sequence (#seq) count in the training and testing sets of two considered datasets.

	Training Set		Testing Set	
	#doc	**#seq**	**#doc**	**#seq**
i2b2/VA 2010	349	30,673	477	45,025
ShARe/CLEF 2013	200	10,171	100	9,273

4.2 Dataset preparation

In this paper, a sample (in both supervised and active learning approaches) corresponds to a sequence of tokens or a sentence. In the i2b2/VA 2010 dataset, sentence boundaries are already identified as each line in a report file corresponds to a sentence. However, this is not the case for the ShARe/CLEF 2013 dataset. We then used the Leaman's sentence segmentation method [14] to derive sentences for this dataset.

After detecting sentence boundaries, the data was encoded into a representation format that clearly indicates the span of a concept within a sentence. The "BIO" format was leveraged to specify the beginning (B), inside (I), and outside (O) of a concept [23].

We employed a typical feature set, including linguistic (part-of-speech tags), orthographical (regular expression patterns), lexical and morphological (suffixes/prefixes and character n-grams), contextual (window of k words), and semantic features. The Medtex system [18] was leveraged to extract semantic features. Specifically, the UMLS and SNOMED CT (SCT) semantic types for each token were used as the domain knowledge. We leveraged this knowledge to distinguish target from non-target semantic types. A non-quantized value was assigned to each semantic type based on the distribution of each semantic type in the training set according to Equation (12).

4.3 Fully Supervised Approach

We use a tuned linear-chain Conditional Random Fields (CRFs) [12] as benchmark supervised ML algorithm as well as our base ML algorithm in the AL framework. CRFs [13] are the state-of-the-art supervised machine learning approach in clinical concept extraction tasks [20, 34].

The posterior probability of $\vec{y}$ given $\vec{x}$ is described by a linear-chain CRFs model with a set of parameters θ:

$$P_\theta(\vec{y}|\vec{x}) = \frac{1}{Z_\theta(\vec{x})}\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, x_i)\right) \quad (16)$$

Here $Z_\theta(\vec{x})$ is the normalization factor. Each f_j is the transition feature function between label state $i-1$ and i on the sequence x at position i. The $\theta = (\lambda_1, \dots, \lambda_m)$ represent the corresponding feature weights.

Our implementations of CRFs for supervised learning (Sup), both baselines, and all active learning (AL) approaches are based on the MALLET toolkit [17].

4.4 Active Learning Settings

Within this paper, we use an incremental, pool-based, active learning framework. In the incremental approach, a model is not trained from scratch at each iteration: instead, its parameters are updated in successive iterations [12]. 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 effectiveness. We use Medtex to extract the required knowledge for each sample $\vec{x}$ (Equation (13)) in the DKI query strategy.

Random sampling (RS) and longest sequence (LS) are two common baselines for analyzing the benefits of the AL framework. RS randomly selects samples; LS selects samples with the longest length (number of tokens).

For AL and the baseline approaches, the initial labeled set is formed by randomly selecting 1% of the training data. The batch size (B) is set to 200 for i2b2/VA 2010 and 100 for ShARe/CLEF 2013 across all experiments, leading to a total of 153 and 101 batches, respectively.

While in real settings AL would use human annotators to label informative samples at each iteration, here we simulate this process by using the annotations provided in the training set of the two datasets.

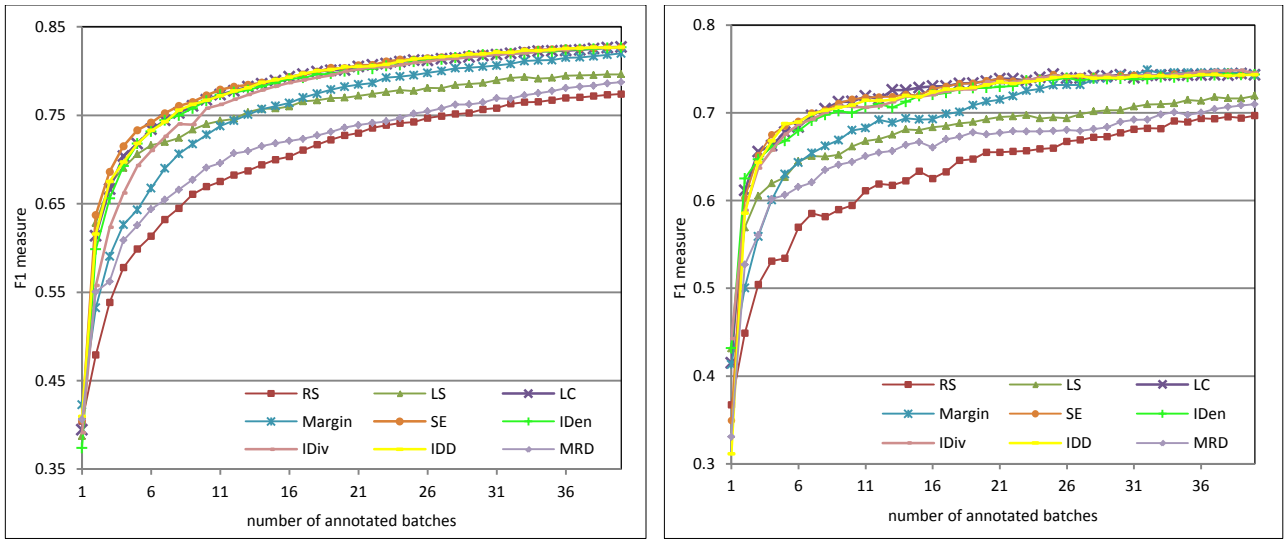
4.5 Evaluation Measures

In our evaluation, concept extraction effectiveness is measured by Precision, Recall and F1-measure. Evaluation metrics are computed on test data using MALLET's multi-segmentation evaluator[4]. Query strategies were analyzed and compared among each other and against the fully supervised method based on their learning curves across batches. Learning curves allow to associate model performance and annotation effort.

A core aspect of our evaluation is determining the first iteration at which AL strategies achieve an effectiveness comparable (i.e., equal or higher) to that of the supervised method. This allows to identify the *minimum amount of annotated data* the AL strategies require to achieve the same effectiveness of the supervised method (target effectiveness): this corresponds to the point of intersection between the AL learning curve and the target effectiveness.

We further analyze the results by considering the Annotation Rate

[4] The results obtained with this evaluator are not directly comparable to those reported using the i2b2 or ShARe/CLEF evaluation scripts.



(a) **(b)**

Figure 3. The effectiveness of active learning approaches and baselines for the first 40 batches (a) i2b2/VA 2010 (b) ShARe/CLEF 2013.

(AR), which measures how much annotation effort is saved by an AL approach. AR is the number of labeled annotation units in terms of sequences (SAR), tokens (TAR) and concepts (CAR) used by AL to reach the target effectiveness, over the number of labeled annotation units used by the supervised method.

$$\text{AR} = \frac{\#\ labeled\ annotation\ units\ used\ by\ AL}{\#\ total\ labeled\ annotation\ units\ used\ by\ Sup} \quad (17)$$

The lower the AR, the less annotation effort is required. Note that here every annotation unit (sequence, token, and concept) is considered as having the same annotation cost (uniform annotation cost). While this setting may not be fully representative of real-world use cases [28], no annotation cost is distributed along with the considered datasets and the literature lacks of specific studies that consider annotation costs for clinical concept extraction.

5. RESULTS

We first study to what extent various active learning query strategies reduce the annotation effort for clinical data compared to a fully supervised approach: this serves to determine the state-of-the-art strategy for clinical concept extraction among those proposed in the literature. We then compare the best results from the state-of-the-art with our proposed query strategy to assess if domain knowledge based query strategies further reduce the burden of manual annotation in clinical settings.

5.1 Target Effectiveness for Active Learning

Table 2 presents the effectiveness of the fully supervised CRF model.

Table 2. Effectiveness of the fully supervised approach (Sup) (P = Precision, R = Recall, F1 = F1-measure).

	i2b2/VA 2010			ShARe/CLEF 2013		
	P	**R**	**F1**	**P**	**R**	**F1**
Sup	0.8409	0.8066	0.8234	0.8095	0.6804	0.7394

These results are used as the target effectiveness for determining the level of annotation effort savings contributed by different active learning approaches.

5.2 Analysis of AL Query Strategies on Clinical Data

5.2.1 Query Strategy Performance

Figure 3 shows the learning curves obtained with different active learning query strategies in the first 40 batches of the i2b2/VA 2010 and ShARe/CLEF 2013 datasets.

For both datasets, the learning curves of all query strategies are always well above the RS baseline, as expected. However, MRD always performs worse than the longest sequence baseline, suggesting that only relying on the similarity between samples to select subsequent batches in the AL loop is not effective. This further highlights that the similarity measure on its own is not enough to find useful samples for active learning. We hypothesize that this is because clinical data is partially characterized by the repetition of fairly similar patterns [11]. Although the diversity element prevents the MRD approach from selecting similar samples, it still fails to pick the most informative ones.

The fact that the learning curves of approaches that leverage informativeness in addition to similarity measures are higher than the learning curve of MRD demonstrates the importance of informativeness in selecting useful samples.

In i2b2/VA 2010, the LS baseline effectiveness is higher than IDiv in the first six batches (F1 = 0.7). The Margin strategy also performs poorly compared to the LS baseline in early batches, but outperforms LS after ten batches in i2b2/VA 2010 (F1 = 0.74) and six batches in ShARe/CLEF 2013 (F1 = 0.63). We hypothesize that this is because, in clinical datasets, longer sequences usually contain more concepts. Hence, by choosing the longest sequences, the effectiveness of LS increases quite sharply in early batches as it already reaches a reasonable effectiveness[5] compared to the

[5] i2b2/VA 2010, F1 ≈ 0.73 and ShARe/CLEF 2013, F1 ≈ 0.63.

Table 3. Annotation rates for all active learning query strategies and baselines.

		SAR(%)		TAR(%)		CAR(%)	
		i2b2/VA 2010	ShARe/CLEF 2013	i2b2/VA 2010	ShARe/CLEF 2013	i2b2/VA 2010	ShARe/CLEF 2013
Baselines	**RS**	90%	97%	90%	97%	90%	97.5%
	LS	58%	72.5%	88%	95%	92%	93.5%
Informativeness-based Approaches	**LC**	24%	24%	43%	38%	55%	63%
	Margin	30%	30%	50%	44%	60%	67.5%
	SE	22%	25%	43.5%	40.5%	54.5%	67%
Informativeness-Similarity Based Approaches	**IDen**	24%	26%	45%	42.5%	55%	67%
	IDiv	23%	24%	41%	38%	52%	63.5%
	IDD	22%	25%	41%	41%	51%	66%
Model-Independent Approaches	**MRD**	68%	74.5%	89%	94%	90%	91%

target effectiveness (Table 2) twice sooner than RS, and it is therefore a very strong baseline. However, after a few iterations the growth rate of the LS learning curve suddenly decreases: this is because the selected long sequences contain no more useful samples to train the model.

5.2.2 Active Learning and Annotation Effort

Annotation rates for all active learning query strategies and baselines are reported in Table 3. The top three approaches are highlighted.

Among the informativeness-based approaches, Least Confidence and Sequence Entropy are very close and outperform Margin.

By adding a similarity element to informativeness, annotation rates in terms of sequences, tokens, and concepts are reduced. However, the informativeness-similarity based approaches exhibit different behaviors across the two datasets. IDiv is the best method in ShARe/CLEF 2013, while IDD performs better than IDiv in i2b2/VA 2010, although differences are not significant.

Since token (TAR) and concept (CAR) annotation rates are more appropriate to measure the actual savings of annotation effort rather than the number of annotated sequences [11], we can conclude that IDD and LC are the most promising approaches for i2b2/VA 2010 and ShARe/CLEF 2013, respectively.

The advantage of LC compared to IDiv and IDD, is that LC is computationally more efficient. Indeed, informativeness-similarity based approaches generally require a considerably larger amount of computations. This is because $\mathcal{R}_{\text{representative}}(\vec{x})$ and $D_{diversity}(\vec{x})$ have to be computed for all samples in the unlabeled and labeled sets, respectively. Although these similarity computations for $\mathcal{R}_{\text{representative}}(\vec{x})$ can be performed prior to starting the active learning loop (as the unlabeled dataset is available upfront in pool-based active learning), $D_{diversity}(\vec{x})$ has to be computed at each iteration of active learning.

5.2.3 Least Confidence vs. Augmented Least Confidence

We now study how iteratively including samples to the labeled set $\mathcal{L}$ for which the model is very confident about their automatically assigned labels (Section 3.1.4) could lead to further annotation effort reduction. To identify these high confidence samples, we experiment with two thresholds, $\tau = 0.998$ and $\tau = 0.999$.

Based on the results reported in Table 4, we can observe that the ALC approach with $\tau = 0.999$ improved the results of LC (Table 3) in both datasets. However, when τ was set to 0.998, in the ShARE/CLEF 2013 dataset ALC required more annotated data to reach the target effectiveness compared to LC.

Table 4. Annotation rates for ALC using two thresholds.

		ALC	
		$\tau = 0.998$	$\tau = 0.999$
i2b2/VA 2010	**SAR(%)**	24.5%	22.5%
	TAR(%)	44%	41.5%
	CAR(%)	55%	52.5%
ShARe/CLEF 2013	**SAR(%)**	25%	22.5%
	TAR(%)	40%	35%
	CAR(%)	67%	58.5%

5.2.4 Initial Labelled Set

While the longest sequence approach (LS) did not reach the target effectiveness much before having trained on all available batches, it is interesting to note that its effectiveness was comparable to the three best approaches (LC, IDD, and IDiv in Table 3) in the first five batches for both datasets. This suggests that long sequences in clinical reports often include useful information to train the classifier, even though its effectiveness increases at a slower rate after the initial batches. Hence, it may be preferable to consider the length of sequences in the early stages of the active learning process so as to select longer sequences first.

A possible way for achieving this is selecting long sequences as the initial labeled set instead of using random selection. Table 5 reports the results for LC, IDiv, IDD, and ALC ($\tau = 0.999$) approaches using the longest sequence approach (LS) for selecting the initial labeled set across both datasets.

Table 5. Annotation rates for LC, IDiv, IDD, and ALC ($\tau = 0.999$) using LS to produce the initial labeled set.

		LC	IDiv	IDD	ALC
i2b2/VA 2010	**SAR(%)**	23%	24%	24%	22.5%
	TAR(%)	42%	41.5%	40%	41%
	CAR(%)	52.5%	52%	53.5%	51%
ShARe/CLEF 2013	**SAR(%)**	26%	25%	22%	24%
	TAR(%)	43%	41%	39%	41.5%
	CAR(%)	67%	65%	60.5%	65%

Table 6. Annotation rates for baselines, benchmarks, and external knowledge based approaches.

		SAR(%)		TAR(%)		CAR(%)	
		i2b2/VA 2010	ShARe/CLEF 2013	i2b2/VA 2010	ShARe/CLEF 2013	i2b2/VA 2010	ShARe/CLEF 2013
Baselines	**RS**	90%	97%	90%	97%	90%	97.5%
	LS	58%	72.5%	88%	95%	92%	93.5%
Benchmarks	**LC**	24%	24%	43%	38%	55%	63%
	IDiv	23%	24%	41%	38%	52%	63.5%
	IDD	22%	25%	41%	41%	51%	66%
External Knowledge-based Approaches	**DKI_t**	22%	28%	39%	32%	51%	63%
	DKI_{tt}	21.5%	20%	27%	31%	37%	57%

Using the longest sequences as the initial labeled set did not show consistent results across datasets and query strategies. In fact, the annotation rates for IDD in ShARe/CLEF 2013, LC and ALC in i2b2/VA 2010 slightly decreased; but other query strategies do not exhibit improvements.

5.3 Domain Knowledge and Active Learning

The CAR results from Section 5.2.2 suggest that at least half of the concepts in the training set are required to be manually annotated before reaching the target effectiveness, even if the most effective AL approaches were used.

In this section, we compare the identified state-of-the-art approaches with the DKI query strategies to investigate how domain knowledge acquired from external resources can help to reduce the annotation effort on clinical data.

The results in Table 6 show that DKI based methods lead to a lower annotation rate compared to baselines and state-of-the-art AL methods. Token-based domain knowledge informativeness (DKI_t, Equation (11)) outperforms most baselines and benchmarks in the i2b2/VA dataset and is as good as IDD (the best approach so far). The results on the ShARe/CLEF 2013 dataset provide however less conclusive findings. If the length of the sequences is used when querying samples with DKI_{tt}, then significant reductions in token and concept annotation rates are achieved on both datasets. In the i2b2/VA 2010 dataset, DKI_{tt} reduces the concept and token annotation rates by 14% compared to IDD (the best state-of-the-art approach for this dataset, see Table 3). In the ShARe/CLEF 2013 dataset, DKI_{tt} reaches the target effectiveness using 6% less annotated concepts and 7% less annotated tokens than LC.

When the performance of DKI_t and DKI_{tt} are compared, it can be observed that the length of sequences has a significant contribution in the annotation effort savings. However, to achieve the best results, sequence length should be considered along with informativeness and domain knowledge, as it does not provide significant AR reductions when considered alone (i.e., LS).

An interesting result is the small difference in sequence annotation rate (SAR) between IDD and DKI_{tt} for the i2b2/VA 2010 dataset. The two approaches select almost the same number of sequences to reach the target effectiveness but with different characteristics (tokens and concepts); this leads to very different results when considering the number of tokens and concepts required to be annotated. This observation demonstrates that domain knowledge combined with a simple informativeness measure (marginal probability) can lead to a better selection of samples; and this is achieved without resorting to using similarity measures which are computationally costlier. DKI_{tt} shows a varied but significant range of savings in terms of required number of concepts to be annotated across the two datasets (37%-57%).

6. RELATED WORK

Active learning aims to achieve high effectiveness (at least comparable to that of supervised methods), while reducing annotation costs by minimizing the amount of data that is required to be annotated. Domains like the clinical one are characterized by high costs for obtaining (expert) annotations. In such domains, it becomes then critical to reduce the annotation effort while ensuring no loss of effectiveness for automated techniques.

One of the main challenges in active learning is identifying and querying samples that can better inform classification models [26]. Different strategies for selecting informative samples have been proposed, including: uncertainty sampling [15], query-by-committee [29], and information density [27]. Settles and Craven [27] performed an extensive empirical evaluation of different query selection strategies within active learning, using different corpora for sequence labeling tasks. They found that information density and sequence vote entropy outperform the state-of-the-art in active learning in most corpora. However, in this paper we found that diversity along with information density lead to better sample selection for concept extraction in the clinical domain. We speculate that this is because of the high similarity between sequences in clinical narratives. Hence, samples selected by information density are not useful for training classification model within the active learning process, while diversity allows to select more representative samples.

While the effectiveness of active learning methods has been conclusively demonstrated in many domains such as text classification, information extraction and speech recognition [26], as highlighted, clinical and biomedical natural language processing tasks have seen only limited use of active learning [19]. To demonstrate the key role of AL in reducing the burden of manual annotation, in previous work we built a preliminary active learning-based system and investigate the state-of-the-art AL methods (LC and IDen) for extracting medical concepts from clinical free text. The annotation effort saved by AL to achieve the target effectiveness (supervised effectiveness) was up to 54% of the total number of concepts (CAR) [11]. We also investigated the factors that influence the robustness and the effectiveness of models learnt within the active learning frameworks [12]. It was found that well selected feature set, the incremental learning setting, and the tuning of the supervised classifier parameters lead to more robust active learning models. Here we consider the same settings and parameters, and observe that the models are generally robust, as demonstrated by the learning curves (Figure 3).

Within the clinical domain, active learning has been used for classifying clinical concepts according to their assertions [3, 22] and co-reference resolution [36]. For assertion classification, Chen, et al. [3] introduced a "model change" sampling-based algorithm and found it performed better than random sampling, uncertainty sampling-based algorithms, and information density-based algorithm. Active learning has also been used for de-identifying Swedish clinical records [2]. There, the information extraction problem was treated as a classification task in which words belong to one of eight personal clinical information types or to the non-personal clinical information type. The most uncertain and the most certain sampling strategies were evaluated using the highest and lowest entropy, and it was found that these methods outperformed a random sampling baseline. Figueroa, et al. [6] analyzed the effectiveness of different active learning methods, including distance-based (DIST), diversity-based (DIV), and a combination of both (CMB), based on clinical data characteristics. However, this study focused on a clinical classification task and only applied a limited number of query strategies. They found that the effectiveness of DIV and DIST is strongly dependent on the dataset diversity and uncertainty, respectively. Rosales, et al. [22] presented a semi-supervised active learning framework based on query by committee and evaluated its effectiveness in a binary classification task, where concepts extracted from clinical free text are classified into two classes: present or absent/negated. Their framework is able to select both informative and representative samples.

7. CONCLUSION AND FUTURE WORK

This paper has introduced a new active learning query strategy, called domain knowledge informativeness (DKI). This novel query strategy leverages domain knowledge resources, such as ontologies and terminologies, for extracting concepts from text. DKI is instantiated and evaluated within the clinical domain for the task of clinical concept extraction. Our instantiation considers the longest span semantic value, as obtained from a clinical NLP tool, and the marginal probability of a text sequence, as assessed by a classification model. These are then used to query samples for selecting those that not only are informative but also contain useful concepts based on the external domain knowledge.

Our empirical evaluation highlights the promise of integrating domain knowledge within active learning query strategies, as indicated by the gains in annotation effort reduction achieved by DKI over state-of-the-art approaches.

We also found that, when analyzing the performance of active learning on clinical data, the confidence about samples' labels estimated by the classification model is a deciding factor in selecting effective samples.

Our study has two main limitations. First, all evaluation metrics used in this study cannot be directly translated into actual cost reduction. However, CAR and TAR values can be used to design a cost model for real world cases. Second, in real world settings the full training data is not available in advance for estimating the target effectiveness, thus making the decision of where to stop the active learning process is a challenging problem, which requires to consider a trade-off between cost and effectiveness.

In order to overcome these limitations, our future work will consider designing a cost model that uses the required number of annotated tokens and concepts. This cost model would constitute a step forward for precisely assessing the actual contribution of active learning in terms of cost reduction. At the same time, we plan to explore methods to determine a stopping point independently from the supervised effectiveness.

8. REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 17-21, 2001.

[2] H. Boström and H. Dalianis. De-identifying health records by means of active learning. *Recall (micro)*, 97(97.55), 90-97, 2012.

[3] Y. Chen, S. Mani, and H. Xu. Applying active learning to assertion classification of concepts in clinical text. *Journal of Biomedical Informatics*, 45(2), 265-272, 2012.

[4] R. A. Cote and S. Robboy. Progress in medical information management. *Journal of the American Medical Association (JAMA)*, 243(8), 756-762, 1980.

[5] A. Culotta and A. McCallum. Reducing labeling effort for structured prediction tasks. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*, 746–751, 2005.

[6] R. L. Figueroa, Q. Zeng-Treitler, L. H. Ngo, S. Goryachev, and E. P. Wiechmann. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association (JAMIA)*, 19(5), 809-816, 2012.

[7] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association (JAMIA)*, 11(5), 392-402, 2004.

[8] H. Gurulingappa. Mining the medical and patent literature to support healthcare and pharmacovigilance, *Ph.D. dissertation*, University of Bonn, Bonn, Germany, 2012.

[9] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association (JAMIA)*, 18(5), 601-606, 2011.

[10] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, and G. Zuccon. Overview of the share/clef ehealth evaluation lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, 172-191, 2014.

[11] M. Kholghi, L. Sitbon, G. Zuccon, and A. Nguyen. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association (JAMIA)*, 2015 [In Press].

[12] M. Kholghi, L. Sitbon, G. Zuccon, and A. Nguyen. Factors influencing robustness and effectiveness of conditional random fields in active learning frameworks. In *Proceedings of the 12th Australasian Data Mining Conference (AusDM 2014) (Vol. 158): Conferences in Research and Practice in Information Technology*, 2014.

[13] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, 282-289, 2001.

[14] R. Leaman, R. Khare, and Z. Lu. NCBI at 2013 ShARe/CLEF eHealth Shared Task: disorder normalization in clinical notes with DNorm. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2013)*, 2013.

[15] D. D. Lewis and J. Catlett. Heterogenous Uncertainty Sampling for Supervised Learning. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 148-156, 1994.

[16] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4), 281-291, 1993.

[17] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[18] A. N. Nguyen, M. J. Lawley, D. P. Hansen, and S. Colquist. A simple pipeline application for identifying and negating SNOMED clinical terminology in free text. In *Proceedings of the Health Informatics Conference (HIC)*, 188-193, 2009.

[19] L. Ohno-Machado, P. Nadkarni, and K. Johnson. Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature. *Journal of the American Medical Informatics Association (JAMIA)*, 20(5), 805, 2013.

[20] S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, and G. Savova. Task 1: ShARe/CLEF ehealth evaluation lab 2013. *CLEF 2013 Evaluation Labs and Workshops: Working Notes*, 2013.

[21] S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, and G. Savova. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association (JAMIA)*, 22(1), 143-154, 2015.

[22] R. Rosales, P. Krishnamurthy, and R. B. Rao. Semi-supervised active learning for modeling medical concepts from free text. In *Proceedings of the Sixth International Conference on Machine Learning and Applications*, 530-536, 2007.

[23] E. F. T. K. Sang and J. Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 173-179, 1999.

[24] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association (JAMIA)*, 17(5), 507-513, 2010.

[25] T. Scheffer, C. Decomain, and S. Wrobel. Active hidden markov models for information extraction. In *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)*, 309-318, 2001.

[26] B. Settles. Active learning, (Vol. 6): Morgan & Claypool Publishers, 2012.

[27] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1070-1079, 2008.

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

[29] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 287-294, 1992.

[30] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656, 1948.

[31] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. South, D. Mowery, G. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, and G. Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (Vol. 8138), 212-231, 2013.

[32] Ö. Uzuner, I. Goldstein, Y. Luo, and I. Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association (JAMIA)*, 15(1), 14-24, 2008.

[33] Ö. Uzuner, I. Solti, and E. Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association (JAMIA)*, 17(5), 514-518, 2010.

[34] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association (JAMIA)*, 18(5), 552-556, 2011.

[35] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association (JAMIA)*, 17(1), 19-24, 2010.

[36] H.-T. Zhang, M.-L. Huang, and X.-Y. Zhu. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6), 1302-1313, 2012.

[37] X. Zhu. Semi-supervised learning literature survey. *Technical Report 1530*: University of Wisconsin-Madison, 2007.