JMLR: Workshop and Conference Proceedings 16 (2011) 169–181 Workshop on Active Learning and Experimental Design

Inspecting Sample Reusability for Active Learning

Katrin Tomanek KATRIN.TOMANEK@UNI-JENA.DE
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany

Katharina Morik KATHARINA.MORIK@TU-DORTMUND.DE
Department of Computer Science - Artificial Intelligence Group
Technical University of Dortmund, Germany

Editor: I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov

Abstract

Active Learning (AL) exploits a learning algorithm to selectively sample examples which are expected to be highly useful for model learning. The resulting sample is governed by a sampling selection bias. While a bias towards useful examples is desirable, there is also a bias towards the learner applied during AL selection. This paper addresses sample reusability, i.e., the question whether and under which conditions samples selected by AL using one learning algorithm are well-suited as training data for another learning algorithm.

Our empirical investigation on general classification problems as well as the natural language processing subtask of Named Entity Recognition shows that many intuitive assumptions on reusability characteristics do not hold. For example, using the same algorithm during AL selection (called selector) and for inducing the final model (called consumer) is not always the optimal choice. We investigate several putatively explanatory factors for sample reusability. One finding is that the suitability of certain selector-consumer pairings cannot be estimated independently of the actual learning problem.

Keywords: active learning, uncertainty sampling, sample selection bias, covariate shift

1. Introduction

While supervised machine learning methods are de-facto standards for a variety of real-world problems, their greediness for large amounts of labeled training data is one of the major obstacles along the path to applications. Training data are usually not available in real-world applications. Human experts of the specific domain in focus need to create such labeled examples which is extremely costly. This holds, for example, for a range of natural language processing (NLP) tasks where (parts of) natural language text need to be classified. For the creation of training data the human annotator has to read through a set of (usually randomly selected) textual examples and manually assigns the corresponding categories to the constituent of interest (e.g., words). Such annotation is costly and usually requires domain experts, for example when biomedical publications are to be annotated.

Active learning (AL) tackles the challenge of economic training data creation. In the AL paradigm, only examples of high utility for classifier training are selected for manual annotation in an iterative manner. AL has been shown to be a promising solution to annotation cost reduction, especially for scenarios where large amounts of unlabeled data are

available at no or relatively low costs. Since AL exploits a learning algorithm to selectively sample examples which are expected to be highly useful for model learning, the resulting sample is governed by a sampling selection bias, also known as covariate shift. While a bias towards useful examples is generally intended and desirable, one must also keep in mind that utility is assessed with respect to the learner applied during AL selection so that the resulting sample is somewhat biased to this particular learner.

Approaches to AL are typically based on the assumption that the learning algorithm used during selection – called *selector* – and the learning algorithm used to induce the final model – called *consumer* – to are identical. Yet, there are settings where selector and consumer intentionally diverge. Firstly, the interaction with the annotating expert demands a fast learning algorithm embedded in AL. Hence, a less complex learner might be used as selector, while the final consumer remains the high-accuracy, more complex learning algorithm. Secondly, the optimal learner for a new problem is often unknown during data acquisition so that the selector is likely to differ from the final consumer. Thirdly, we may want to annotate just once and use the example set for many different learning problems. We call settings, where selector and consumer diverge *foreign-selection* (in contrast to *self-selection* as the default setting). Foreign-selection constitutes a scenario of AL *sample reuse*. We say that a sample S_{AL} obtained by AL is *reusable* by a particular consumer, if this consumer yields a higher classification accuracy when trained on S_{AL} than it would (on average) achieve when trained on a random sample S_{RD}.

The question is, whether and under which conditions a sample selected with AL exploiting a specific learner is suitable (i.e., reusable) for training another learning algorithm, or – more generally – how sampling efficiency of AL is affected by foreign-selection settings. To the best of our knowledge, this question has neither been posed nor studied in context of AL before. Most research in AL is restricted to a self-selection scenario. Despite its practical importance, the reusability issue has not yet been investigated.

For our investigation of reusability we state a set of hypotheses on a) expected reusability characteristics of specific foreign-selection scenarios and b) relevant factors assumed to influence reusability in foreign-selection scenarios. These hypotheses are empirically tested on several general classification problems as available from the UCI repository as well as the NLP task of Named Entity Recognition (NER) which is a well acknowledged prerequisite for tailored information services and so an inherently realistic application scenario of AL (Tomanek et al., 2007).

The rest of this paper is structured as follows: Section 2 motivates the hypotheses we aim to test. Section 3 then describes our experimental setting, including data sets, learning algorithms, AL approaches, and a novel measure for reusability. Results are reported and discussed in Section 4. Related work is discussed in Section 5 and Section 6 concludes.

2. Hypotheses

Expected reusability characteristics The first two hypotheses state what seems to be common-sense:

- **H1:** Samples obtained by AL with a particular selector are rather *unlikely to be reusable* by another learning algorithm due to adversarial ties to the selector.

- **H2:** For a particular consumer, self-selection constitutes the *upper bound* for AL sampling efficiency.

Expected factors influencing reusability These hypotheses cover four factors which possibly influence the reusability characteristics of a specific scenario.

- **H3:** Are there selector-consumer pairings exhibiting *general reusability characteristics* which hold for most learning problems? H3 states, that there are selectors which are in general well suited for certain consumers, and vice-versa.
- **H4:** Since the selector classifies examples according to its model, the similarity of the consumer's and the selector's model could determine reusability. H4 states that a high degree of *model similarity or model relatedness* leads to high reusability.
- **H5:** Since the resulting selection is what counts, H5 states that the *similarity of samples* chosen by self-selection and foreign-selection is important for reusability.
- **H6:** Since the example input space is changed by a learner's feature weights, H6 states that the *similarity of the feature ranking* in self- and foreign-selection is important for reusability.

3. Experimental Setup

This section outlines the experimental setup used for empirically investigating the hypotheses on AL sample reuse.

3.1. Learning Problems and Data

General classification problems as well as the NER learning problem are chosen. The data sets are chosen such that results are reproducible and comparable to previous work. Within the UCI repository (Asuncion and Newman, 2007), five data sets were selected according to *(a)* size (data sets should have more than 1,000 examples so that AL can actually select), and *(b)* diversity (data sets should contribute different numbers of features and example/feature ratios as well as different numbers of target classes). As for NER, we chose the Muc7 and the PBgene corpus. Both corpora consist of natural language sentences annotated with respect to the particular entity classes of interest. [1] Table 1 gives an overview of the selected data sets and corpora.

3.2. Learning Algorithms

For experiments on the UCI data sets, we chose four well-known learning algorithms: Naïve Bayes (NB), Multinomial Logistic Regression (MaxEnt), C4.5 Decision Trees (DT), and linear kernel Support Vector Machines (SVM). The respective implementations of these algorithms in the WEKA toolkit are used with their default parameters (Witten and Frank, 2005). For NER experiments, we applied the following algorithms: Conditional Random Fields (CRF), MaxEnt, NB, Hidden Markov Models (HMM), and SVMs. We used standard features for NER (Nadeau and Sekine, 2007).

1. Muc7 (see http://www.ldc.upenn.edu/Catalog) has 7 entity types; PBgene is a sub-corpus derived from the PennBioIE corpus (see http://bioie.ldc.upenn.edu/) and has 3 gene entity types.

UCI data sets

data set	# examples	# attributes	attribute types	classes
CAR	1,728	6	nominal	4
MUSHROOM	8,124	22	nominal	2
NURSERY	12,960	8	nominal	5
SEGMENT	2,310	19	real	7
SICK	3,772	30	mixed	2

NER data sets

data set	# examples	# attributes	attribute types	classes
MUC7	3,022	$\approx$ 50K	binary	8
PBGENE	10,570	$\approx$ 50K	binary	4

Table 1: Data sets used for AL sample reuse experiments. For NER, *examples* refers to the number of sentences contained in the respective data set.

3.3. Active Learning and Utility Measures

In the scenario inspected here, the expert is in the loop of AL. This requires a fast processing of AL. Fast utility estimates come along with the price of possibly not finding a globally optimal sample. Statistically optimal approaches to AL (such as in Cohn et al. (1996) or Roy and McCallum (2001)) usually require model retraining for each unlabeled example to be tested in each AL iteration. In contrast, Uncertainty Sampling (Lewis and Gale, 1994) requires only one model training step in each AL iteration. Uncertainty Sampling correlates utility with model confidence: the utility of an example is based on the uncertainty (as the inverse of the confidence) of the current classifier in its prediction. For our experiments we thus decided for Uncertainty Sampling instead of statistically optimal approaches to fit the practical requirement of low selection times when an (annotation) expert is in the loop.

In each iteration, AL greedily selects example $p = (x)$ with the highest utility score $u(p, \theta)$ which is based on the current model θ. Such a locally optimal selection depending on the history of previous selections is performed with the hope that it will lead to a good global solution. The true class label y for a selected example is queried from a human expert and the labeled example is then added to the training set $\mathcal{L}$ and the next AL iteration starts. After stopping, the sample $\mathcal{L}^*$ containing all labeled examples is returned. This sample is then used to train the final model. Algorithm 1 formalizes this procedure. When applied on the UCI data sets, we indeed only selected one example per AL iteration. Applied to the more complex learning problem of NER, we modify Algorithm 1 so that $b > 1$ examples with the highest utility scores are selected. This aims at keeping selection time low.

For AL with a NB and a MaxEnt-based selector, the confidence is estimated as the margin between the best and the second best label. The margin utility function (Scheffer and Wrobel, 2001) is given by:

$$u_{\mathrm{MA}}(p, \theta) = 1 - \left(\max_{y' \in \mathcal{Y}} P_\theta(y'|x) - \max_{\substack{y'' \in \mathcal{Y} \\ y' \neq y''}} P_\theta(y''|x) \right) \tag{1}$$

For maximum margin classification, the decision value $d(\mathbf{x}) = \langle \boldsymbol{w}, \mathbf{x} \rangle + b$ indicates the distance of an example to the hyperplane. Larger distances can be interpreted as higher

Algorithm 1: Greedy Active Learning

input :
$\mathcal{L}$: set of labeled examples $l = (x, y) \in \mathcal{X} \times \mathcal{Y}$;
$\mathcal{P}$: set of unlabeled examples $p = (x) \in \mathcal{X}$;
$T(\mathcal{L})$: a learning algorithm;
$u(p, \theta)$: utility function;

repeat
learn model: $\theta = T(\mathcal{L})$;
select $p^* = \text{argmax}_{p' \in \mathcal{P}} u(p', \theta)$;
query label y for p^*: $l^* = (x, y)$;
$\mathcal{L} = \mathcal{L} \cup l^*$, $\mathcal{P} = \mathcal{P} \setminus p^*$;
until *stopping criterion met*;
return $\mathcal{L}^* = \mathcal{L}$

confidence of the classifier in its classification. For the SVM-based selector, the margin utility function is accordingly defined

$$u_{\text{SVM}}(p, \boldsymbol{w}, b) = -\big(d_{y^*}(\mathbf{x}) - d_{y^{**}}(\mathbf{x})\big) \tag{2}$$

with $y^* = \text{argmax}_{y' \in \mathcal{Y}} d_{y'}(\mathbf{x})$ and $y^{**} = \text{argmax}_{\substack{y'' \in \mathcal{Y} \\ y' \neq y''}} d_{y''}(\mathbf{x})$. Due to the well-known instability of decision trees, Uncertainty Sampling should not be applied (Dwyer and Holte, 2007). Instead, a variant of AL known as *Query-by-Committee* (Seung et al., 1992) is promising. The utility of an example is derived from the disagreement within a committee of classifier models $\mathcal{C} = (\theta_1, \ldots, \theta_c)$. In the experiments, committees with $|\mathcal{C}| = 3$ and member θ_i are trained on a subsample $\mathcal{L}'$ of the available training data $\mathcal{L}$ with $|\mathcal{L}'| = \frac{|\mathcal{C}|-1}{|\mathcal{C}|}|\mathcal{L}|$. The Vote Entropy utility function quantifies disagreement (Engelson and Dagan, 1996)

$$u_{\text{VE}}(p, \mathcal{C}) = -\sum_{y' \in \mathcal{Y}} \frac{V(y', x)}{|\mathcal{C}|} \log \frac{V(y', x)}{|\mathcal{C}|} \tag{3}$$

where $V(y', x)$ is the number of committee members θ_i predicting class y'.

As for the NER learning problems, the example grain size is set to complete sentences. However, since sentences consist of single tokens, we calculate the utility scores for each token separately and then average over all tokens of a sentence to get the sentence-level utility score.[2] Moreover, for the NER learning problems, we apply batch-mode AL where $b = 20$ sentences are selected in each AL iteration. Batch-mode AL is here applied to reduce computational complexity of AL because model training for NER is rather complex due to the high-dimensional feature space ($\approx$ 50,000 in this case).

In all experiments, the data sets described above where each split into a pool of AL selected (90%), and a held-out test set (10%) used to calculate learning curves. The results

2. Note that while CRFs and HMMs are actually used to model the sentence as a sequence of tokens $\boldsymbol{x} = (x_1, \ldots, x_n)$, we still calculated the utility score as an average over all token-level utility scores. The token-level score is based on the marginal probability at position i for a sequence $\boldsymbol{x}$. See e.g. (Tomanek and Hahn, 2009) for details.

reported in the following are averages over 20 independent AL runs. For each run, another random split was generated. All experiments are based on the same 20 splits. AL runs were stopped once $|\mathcal{L}| = 150$ was reached (UCI), or once $\mathcal{L}$ consisted of 50,000 tokens (NER).

3.4. Quantification of Sample Reusability

To quantify sample reusability on a continuous scale we introduce a novel measure based on the Area Under the learning Curve (AUC). For a baseline sampling scenario S_{base} (usually random sampling), the learning curve of AL self-selection S_{self}, and that of AL foreign-selection S_{frgn}, the REU score is given by

$$\text{REU}(S_{\text{frgn}}, S_{\text{self}}, S_{\text{base}}, a, b) = \frac{\text{AUC}(S_{\text{frgn}}, a, b) - \text{AUC}(S_{\text{base}}, a, b)}{\text{AUC}(S_{\text{self}}, a, b) - \text{AUC}(S_{\text{base}}, a, b)} - 1 \tag{4}$$

on an interval $[a, b]$ in the learning curve. This score indicates the percentage decrease of AL self-selection sampling efficiency by foreign-selection relative to the baseline sampling scenario, when compared to self-selection. If REU $= 0$, foreign- and self-selection are equally efficient and in the case of REU > 0, foreign-selection would be even better than self-selection. A negative score with $-1 \ll \text{REU} < 0$ indicates that reusability is in evidence but foreign-selection is less efficient than self-selection. Further, we say that reusability is "high" for negative REU scores close to 0, "low" for negative REU scores beyond or just slightly above -1. If REU < -1 we say that reusability is not given since foreign-selection is worse than random selection. Figure 1 visualizes the calculation of the REU score.

Note that a learning curve shows model performance as a function of data acquisition cost. Data acquisition cost may be application-specific. Obviously, the interval $[a, b]$ of the REU score must be chosen on the appropriate unit. As for the UCI data sets, we assume a unit cost per example and set the interval for REU score calculation to $[50; 150]$ so as to exclude the "start-up" phase of AL where the learning curves are naturally very steep as well as to exclude the "convergence-phase" where learning curves usually flatten out. As for the NER scenario where complete sentences are selected, we say that the cost is a function of the sentence length (measured in number of tokens contained). We set the interval for REU score calculation to $[10000; 30000]$ – again, in order to exclude start-up and convergence phase.

4. Results

The REU scores shown in Tables 2 and 3 are used to discuss the hypotheses.

4.1. Hypothesis H1

As for NER, reusability can be recorded for all AL foreign-selection scenarios and REU scores rarely fall below -0.5, indicating a high degree of reusability for this special learning problem. The only exception is that of foreign-selection for a NB-based consumer on the PBgene corpus. On the UCI data sets, in only 15 out of 60 foreign-selection scenarios, sample efficiency considerably drops below that of random selection with REU scores ≤ -1. Moreover, there are only two cases (both on the segment data set) where a sample actively selected by a specific selector is not reusable in by another consumer.

CAR	consumer			
selector	DT	MaxEnt	NB	SVM
DT	0.00	−0.30	−0.47	−0.42
MaxEnt	−0.84	0.00	−1.04	−0.27
NB	−0.84	−0.11	0.00	−0.28
SVM	−0.90	0.26	−0.86	0.00

MUSHROOM	consumer			
selector	DT	MaxEnt	NB	SVM
DT	0.00	−0.59	−0.69	−0.19
MaxEnt	−0.05	0.00	−0.47	0.12
NB	0.00	−0.07	0.00	−0.02
SVM	−0.91	−0.43	−1.17	0.00

NURSERY	consumer			
selector	DT	MaxEnt	NB	SVM
DT	0.00	−1.13	−2.13	−0.93
MaxEnt	−0.33	0.00	−1.46	−0.07
NB	0.48	−0.09	0.00	0.14
SVM	−0.36	−0.35	−1.19	0.00

SEGMENT	consumer			
selector	DT	MaxEnt	NB	SVM
DT	0.00	−0.72	−0.47	−0.95
MaxEnt	−0.95	0.00	−1.24	−3.07
NB	−2.56	−2.04	0.00	−1.77
SVM	−4.35	−3.53	−2.39	0.00

SICK	consumer			
selector	DT	MaxEnt	NB	SVM
DT	0.00	−0.83	−0.54	−0.90
MaxEnt	0.26	0.00	−0.43	−0.77
NB	0.31	−0.90	0.00	−2.18
SVM	0.34	−0.18	−0.25	0.00

Table 2: Reusability scores (REU) on UCI data sets. Colors: REU ≥ 0 and REU ≤ -1

MUC7 (NER)

	consumer				
selector	NB	HMM	MaxEnt	SVM	CRF
NB	0.00	0.07	−0.19	0.13	−0.15
HMM	−0.48	0.00	−0.40	−0.29	−0.39
MaxEnt	−0.39	−0.05	0.00	0.12	−0.12
SVM	−0.40	−0.07	−0.20	0.00	−0.24
CRF	−0.38	0.01	0.02	0.05	0.00

PBGENE (NER)

	consumer				
selector	NB	HMM	MaxEnt	SVM	CRF
NB	0.00	−0.02	−0.01	−0.17	−0.13
HMM	−1.51	0.00	−0.29	−0.38	−0.35
MaxEnt	−3.47	−0.57	0.00	−0.08	−0.09
SVM	−2.22	−0.24	−0.22	0.00	−0.25
CRF	−3.58	−0.58	−0.06	−0.36	0.00

Table 3: Reusability scores on NER data sets. Colors: REU ≥ 0 and REU ≤ -1

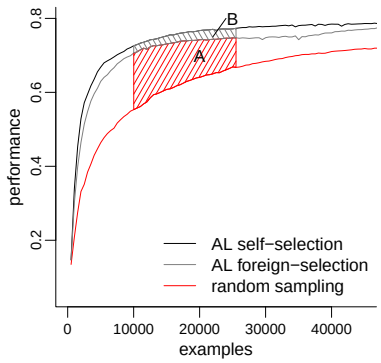

Figure 1: Quantification of sample reusability through the REU score which is here calculated by $\frac{A}{A+B} - 1$. In this example, REU $= -0.17$ indicates good reusability.

The hypothesis that reusability were a rare scenario (H1) is thus rejected. In contrast, reusability is observed in the majority of cases.

4.2. Hypothesis H2

Most surprising, the results reveal that self-selection sampling efficiency is occasionally outperformed by foreign-selection. As for NER, this is the case in 6 out of the 40 foreign-selection scenarios; for the UCI data sets, 8 out of the 60 foreign-selection scenarios exceed the assumed upper bound. Look, for instance, at the combination of an SVM selector and a MaxEnt consumer processing the CAR data set. This falsifies the upper-bound hypothesis leading to the remarkable finding that there are scenarios where a learner T_2 estimates the utility of an example for learner T_1 more appropriately than T_1 itself.

4.3. Hypothesis H3

The shown REU scores also contradict the assumption that there are certain pairings of learning algorithms for which general reusability characteristics hold. Inconsistent reusability characteristics for the selector-consumer pairings have to be ascertained over the five UCI data sets. NB, for example, is a good selector for the MaxEnt consumer on some data sets (MUSHROOM, NURSERY and CAR), but not on others (SICK and SEGMENT).

4.4. Hypothesis H4

Additional experiments test whether model similarity explains reusability. In line with Baldridge and Osborne (2004), H4 states that sample reusability depends on the degree of relatedness between selector and consumer regarding their *models*. Relatedness is usually measured by the degree of correlation between the *predictions* of models. However, in the context of AL, more interesting than the consistency of predictions is how similar the utility rankings of the unlabeled examples achieved with two different models are. This is because these rankings determine which examples are selected. For use in our AL scenario, we thus say that two models are related when they lead to a highly correlated utility ranking of unlabeled examples.

The Spearman's rank correlation coefficient r_S compares the utility rankings of two models (Baldridge and Osborne, 2004). We trained all learners on random samples (of 10,000 tokens in case of NER, and on 150 examples in case of the UCI data sets). [3] Utility rankings of the examples in the test set are then compared for all tuples of models.

Hypothesis H4 is operationalized by the assumption that relatedness scores positively correlate with reusability scores. Table 4 shows correlation coefficients between the REU score and model similarity (relatedness of models). It indicates that there is either no or even a negative correlation between reusability and model similarity. By definition, relatedness scores are symmetrical. However, reusability, according to Table 2, is not. When the learners of selector and consumer are exchanged, reuse scores differ. Looking at the relatedness scores (omitted here due to space limitations), we observe many such cases. As an example, consider the pair of a SVM and a MaxEnt learner for which we obtained a high relatedness score on the PBGENE corpus. A sample obtained by AL with a MaxEnt-based selector is perfectly reusable by an SVM-based consumer. However, when SVM is used to select for a MaxEnt consumer, reusability drops to a REU score of -0.22 (cf. Table 2). As another example, the very good reusability of a sample obtained by AL with a NB selector for a HMM consumer is in contrast to the rather low relatedness score for HMM and NB on the MUC7 corpus. A low relatedness score thus does not necessarily imply a low level of reusability. While a high rank correlation coefficient often accompanies reusability (as for the MaxEnt-CRF tuple), one cannot conclude the opposite from low correlation coefficients. This emphasizes that different samples can also lead to similar model performances.

4.5. Hypothesis H5

H5 hypothesizes that the similarity of samples obtained in self- and foreign-selection mode is a relevant factor for reusability. Different selectors may select from other parts of the instance space. The more the covered space of a foreign-selector diverges from that of the self-selector, the lower the REU scores are according to our assumption. A situation with REU ≤ -1 would then mean that the AL sample does not cover the relevant areas for the consumer. Comparing the sample distributions over the input space is performed by agglomerativ clustering over all unlabeled examples in the pool $\mathcal{P}$. The distance between two clusters is calculated according to the average linkage method based on the Euclidean distance. The hierarchical clustering is flattened down to $k = 20$ clusters. The examples in a sample S are then assigned to clusters in this clustering according to an example's proximity to a cluster centroid (Everitt et al., 2001).

D_S represents the distributions of the examples of sample S over the clustered input space. This distribution gives the percentage of a sample's examples falling in each cluster. Figure 2 visualizes this for two samples S_1 and S_2 obtained by two different selectors. The similarity of the two samples S_1 and S_2 is estimated based on the divergence of their distributions D_{S_1} and D_{S_2} which is calculated by the Jensen-Shannon divergence (JSD). The JSD score ranges in the interval of $[0, 1]$; lower scores indicate higher distributional similarity. The similarity is calculated by $\mathrm{SIM}(S_1, S_2) = 1 - \mathrm{JSD}(D_{S_1}, D_{S_2})$. In the above example, a similarity of $\mathrm{SIM}(S_1, S_2) = 0.48$ is obtained.

3. We also tested random samples of different sizes but did not obtain essentially different results.

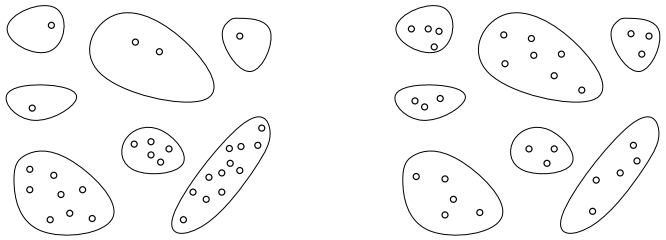

Figure 2: Distribution of samples S_1 (left) and S_2 (right) over common clustering.

correlation of reusability and model similarity

	CAR	MUSHROOM	NURSERY	SEGMENT	SICK
r_P	0.04	-0.31	0.07	-0.22	-0.47
r_S	0.02	-0.31	0.01	-0.39	-0.24

correlation of reusability and sample similarity

	CAR	MUSHROOM	NURSERY	SEGMENT	SICK
r_P	0.30	0.24	0.03	0.37	0.23
r_S	0.29	0.19	0.08	0.40	0.14

correlation of reusability and feature ranking similarity

	CAR	MUSHROOM	NURSERY	SEGMENT	SICK
r_P	0.16	-0.35	0.46	-0.38	-0.06
r_S	0.04	-0.49	0.45	-0.42	-0.21

Table 4: Pearson's (r_P) and Spearman's (r_S) correlation coefficients for REU score and other variables (model similarity, sample similarity, and feature ranking similarity).

Now, H5 becomes the testable statement that similarity (SIM) correlates with reusability (REU). Table 4 shows correlation coefficients between reusability and sample similarity. Pearson's correlation coefficients range between 0.03 and 0.37, which indicates a comparatively low (linear) relationship. Spearman's correlation coefficients are also very low on average ranging from 0.08 to 0.4. SIM scores are symmetrical, where reusability is not. These experiments show that the distributional similarity of samples does not explain reusability.

4.6. Hypothesis H6

Hypothesis H6 states that feature weighting is a relevant factor for reusability. This sensitivity can be expressed by comparing feature rankings obtained from foreign-selection and from self-selection. Feature rankings are obtained by a wrapper approach based on simple hill climbing (Kohavi and John, 1997). Subsequently, tuples of feature rankings are compared. A tuple always consists of the feature ranking obtained from a model learned on a foreign-selection sample and the feature ranking of a model learned on the self-selected sample. Comparison of feature rankings is based on a weighted version of Spearman's rank

correlation coefficient. [4] Accordingly, the *feature ranking score* $\mathrm{FR}(S_{T_1}, S_{T_2})$ shows the correlation of the feature rankings of a model induced by learner T_2 on a foreign-selection sample from AL with a selector based on T_1 and a self-selection sample where the selector was based on T_2.

Now, H6 means that the FR scores correlate highly with the REU scores in the foreign-selection scenarios. However, the experiments disprove this assumption. Table 4 shows correlation coefficients between reusability and feature ranking similarity (FR score). Correlation coefficients are mostly low or even negative. Overall, this outcome shows that the FR score is inadequate for predicting the REU score (Pearson's coefficient) as well as for ranking the selectors according to their appropriateness for a particular consumer (Spearman's coefficient). A twisted feature ranking may still lead to a model with similar accuracy compared to a model which is induced from a self-selected sample.

Reusability cannot be explained by the fact that models learned on different samples exhibit similar feature rankings. Note, a model θ induced from a foreign-selection sample may perform similarly well or even better than a model θ' induced by the same learner but from a self-selection sample.

5. Previous Work

While there is a huge body of work on AL for the self-selection scenario (see Settles (2009) for an overview), there is only little on scenarios of AL sample reuse and foreign-selection. A scenario of sample reuse motivated by the need to reduce the computational complexity of sampling was first described by Lewis and Catlett (1994) for a text classification problem. There, the consumer was based on decision trees and the selector was a logistic regression algorithm. Positive findings about sample reusability were reported. For the NLP task of statistical parsing, controversial findings on reusability have been published. Hwa (2001) reported positive results, Baldridge and Osborne (2004), on the other hand, presented and discussed scenarios where the AL foreign-selection bias considerably impairs reusability.

Previous work on AL sample reuse addresses AL sample reuse only in very specific scenarios. There is to-date no comprehensive study of the true nature of reusability, requirements for the presence of reusability, or prohibitive factors. To the best of our knowledge, this paper is the first approach in this direction.

Under a more general consideration and without explicit reference to AL, Fan et al. (2005) study how sensitive learning algorithms are in general to sample selection bias. A learner is called *local* if it is invariant to this bias, and *global* otherwise. They found that all learning algorithms can be both local and global depending on the combination of the data set, modeling assumptions made by the learner, and the learner's appropriateness to model the particular data set. This observation fits to the findings presented in this paper.

6. Summary and Conclusions

This paper describes the problem of sample reusability in context of AL foreign-selection scenarios. Several hypotheses on reusability characteristics and explanatory factors for

4. A weighted rank correlation is simply calculated based on weighted covariance. The weights are assigned inverse to their ranks.

reusability are empirically investigated. Experiments were performed both on general classification problems and on the NER task which constitutes a special class of learning problems.

Based on the results of our experiments we have to reject the dominant self-selection assumptions (H1, H2). In particular for the NER learning problem, reusability is evident in all practical scenarios. Self-selection does not constitute the upper bound of sampling efficiency but can sometimes be outperformed by foreign-selection. None of the assumed influencing factors – *viz.* model similarity, sample similarity, and similarity of the feature ranking – were supported by our experiments. Reusability could even be observed when all these assumptions were violated. Most importantly, experiments showed that one cannot generalize which combinations of learners generally work well together in settings of AL foreign-selection. Hence, whether reusability is in evidence for a particular selector-consumer pairing appears to depend on the combination of learning problem, data set, and appropriateness of the particular learning algorithm.

Overall, our study points out that reusability is a relevant and challenging problem. Future work in this direction should focus on the quantification of a learner's sensitivity to sample selection bias given a specific learning problem in order to estimate – ideally prior to AL sample selection – whether a sample obtained by AL and a specific selector may be reusable by (which?) consumers. Our measure to quantify sample reusability is ready to use for such further investigations.

References

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

Jason Baldridge and Miles Osborne. Active learning and the total cost of annotation. In *Proc. of EMNLP'04*, pages 9–16, 2004.

David Cohn, Zoubin Ghahramani, and Michael Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

Kenneth Dwyer and Robert Holte. Decision tree instability and active learning. In *Proc. of ECML'07*, pages 128–139, 2007.

Sean Engelson and Ido Dagan. Minimizing manual annotation cost in supervised training from corpora. In *Proc. of ACL'96*, pages 319–326, 1996.

Brian Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Wiley, 4th edition, 2001.

Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proc. of ICDM'05*, pages 605–608, 2005.

Rebecca Hwa. On minimizing training corpus for parser acquisition. In *Proc. of ConLL'01*, pages 1–6, 2001.

Ron Kohavi and George John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

David Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proc. of ICML'94*, pages 148–156, 1994.

David Lewis and William Gale. A sequential algorithm for training text classifiers. In *Proc. of SIGIR'94*, pages 3–12, 1994.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of ICML'01*, pages 441–448, 2001.

Tobias Scheffer and Stefan Wrobel. Active learning of partially hidden markov models. In *Proc. of the ECML/PKDD Workshop on Instance Selection*, 2001.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. of COLT'92*, pages 287–294, 1992.

Katrin Tomanek and Udo Hahn. Semi-supervised active learning for sequence labeling. In *Proc. of ACL/IJCNLP'09*, pages 1039–1047, 2009.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *Proc. of EMNLP-CoNLL'07*, pages 486–495, 2007.

Ian Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, 2005.