Active Learning in Noisy Conditions for Spoken Language Understanding

Hossein Hadian
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
hadian@ce.sharif.edu

Hossein Sameti
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
sameti@sharif.edu

Abstract

Active learning has proved effective in many fields of natural language processing. However, in the field of spoken language understanding which is always dealing with noise, no complete comparison between different active learning methods has been done. This paper compares the best known active learning methods in noisy conditions for spoken language understanding. Additionally a new method based on Fisher information named as Weighted Gradient Uncertainty (WGU) is proposed. Furthermore, Strict Local Density (SLD) method is proposed based on a new concept of *local density* and a new technique of utilizing information density measures. Results demonstrate that both proposed methods outperform the best performance of the previous methods in noisy and noise-free conditions with SLD being superior to WGU slightly.

1 Introduction

Spoken language understanding (SLU) is currently an emerging field in the intersection of speech processing and natural language processing (Tur and De Mori, 2011). The task of an SLU system is to extract meaning from speech utterances. Example real-world applications are AT&T's How May I Help You? and BBN's Call Director. In the field of SLU, as well as other fields of natural language processing, gathering data is fairly cheap but labeling is quite expensive and time-consuming. Thus, active learning methods apply very well and can greatly reduce costs. This article evaluates different techniques of active learning in the context of statistical SLU to reduce the labeling effort as much as possible. Also, SLU deals with the most amount of noise, in comparison with other fields of NLP, making robustness one of its most important issues (Tur and De Mori, 2011). Therefore, in this article noisy conditions of SLU are explored too. In this paper, we concentrate on statistical approaches for modeling the SLU system. Specifically conditional random fields (Lafferty et al., 2001) are used with a flat semantic frame to represent meaning and to model the SLU system.

While there have been a couple of studies on active learning in the context of SLU, they have mostly used only methods in the frameworks of uncertainty sampling (Tur et al., 2003; Jars and Panaget, 2008) and query-by-committee (Gotab et al., 2009). In addition, noisy conditions which are an important aspect of SLU have not been addressed thoroughly.

In this paper, performance of various known active learning methods namely uncertainty sampling, query-by-committee, Fisher information ratio (Settles and Craven, 2008) and instability sampling (Zhu

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1081–1090, Dublin, Ireland, August 23-29 2014.

and Ma, 2012) are examined and analyzed in noise-free and noisy conditions of SLU. Also a new method for measuring informativeness of instances based on the Fisher information framework is developed and evaluated along with other methods. Besides, to deal with noisy conditions, the new concept of *local density* and a new technique to utilize density measures are introduced and described.

The rest of this paper is organized as follows: Section 2 briefly describes CRFs, pool-based active learning framework, and selected active learning methods applicable to CRFs. Section 3 describes the first proposed method: Weighted Gradient Uncertainty. Section 4 introduces the local density concept, describes its motives and the proposed method of SLD is described. In Section 5, the noise model is described and experiments are performed in both noisy and noise-free conditions. Finally in Section 6 conclusions are derived.

2 Active Learning and CRFs

CRFs (Lafferty et al., 2001) are statistical graphical models which have demonstrated state-of-the-art accuracy in many fields as well as in SLU. A linear-chain CRF with parameter vector $\vec{\theta}$, defines the probability of $\vec{y}$ being the true label sequence for observation sequence $\vec{x}$ (with length T) as:

$$P(\vec{y}|\vec{x};\vec{\theta}) = \frac{1}{Z_{\vec{\theta}}(\vec{x})} \cdot \exp\left(\sum_{j=1}^{T}\sum_{i=1}^{K}\theta_i f_i(y_{j-1}, y_j, \vec{x}, j)\right). \quad (1)$$

$Z_{\vec{\theta}}(\vec{x})$ is the normalization factor and ensures that sum of $P(\vec{y}|\vec{x};\vec{\theta})$ over all possible labelings equals 1. There are K feature functions $f_k(y_{j-1}, y_j, \vec{x}, j)$ in a linear-chain CRF along with their weights θ_k. Each feature f_k, is a function of the whole observation sequence, the position of current observation and the current and previous labels. Training is the process of finding the optimum weight vector $\vec{\theta}$ to maximize the conditional log-likelihood of training instances in the labeled data set $\mathcal{L}$:

$$\ell(\mathcal{L};\vec{\theta}) = \sum_{(\vec{x},\vec{y})\epsilon\mathcal{L}} \log P(\vec{y}|\vec{x};\vec{\theta}) - \sum_{k=1}^{K}\frac{\theta_k^2}{\sigma^2}. \quad (2)$$

The second term is a regularization penalty to prevent over-fitting. After training, the labels can be predicted using the Viterbi algorithm.

Given: Labeled set $\mathcal{L}$, unlabeled pool $\mathcal{U}$, query startegy $\phi(\cdot)$, query batch size $\mathcal{B}$
repeat
  // learn a model using the current $\mathcal{L}$
  $\theta = train(\mathcal{L})$;
  for $b = 1$ **to** $\mathcal{B}$ **do**
    // query the most informative instance
    $\mathbf{x}_b^* = \underset{x\epsilon\mathcal{U}}{\operatorname{argmax}}\, \phi(\boldsymbol{x})$;
    // move the labeled query from $\mathcal{U}$ *to* $\mathcal{L}$
    $\mathcal{L} = \mathcal{L} \cup \langle \mathbf{x}_b^*, label(\mathbf{x}_b^*)\rangle$;
    $\mathcal{U} = \mathcal{U} - \mathbf{x}_b^*$;
  end
until *some stopping criterion;*

Figure 1. Pool-based active learning (Settles and Craven, 2008).

The focus of this paper is on pool-based active learning in which a learner should select most informative instances for labeling from a pool of unlabeled ones. We adopt the same notation used by Settles and Craven (2008) for the generic pool-based algorithm, sketched in Figure 1. Query strategy

$\phi(\cdot)$ is a function which evaluates how informative an unlabeled instance is. Most methods of active learning are a definition for this function. In the following subsections the best known active learning methods are briefly described.

2.1 Uncertainty Sampling

In this very common framework the learner queries the instance that it is most uncertain how to label. Two methods in this framework proved effective according to Settles and Craven (2008) which are presented here. First is the **least confident (LC)** method:

$$\phi^{LC}(\vec{x}) = 1 - P(\vec{y}^*|\vec{x};\theta), \quad (3)$$

where $\vec{y}^*$ is the most likely label sequence. Second query strategy is the **sequence entropy (SE)** method which measures informativeness of an instance based on entropy in different labelings:

$$\phi^{SE}(\vec{x}) = -\sum_{\vec{y}\epsilon\mathcal{Y}} P(\vec{y}|\vec{x};\theta)\log P(\vec{y}|\vec{x};\theta), \quad (4)$$

where $\mathcal{Y}$ is the set of all possible labelings for $\vec{x}$.

2.2 Query-By-Committee

Query-by-committee (QBC) is another well-studied and common framework for active learning. There are many approaches in this framework, but we use the approach suggested by Settles and Craven (2008) which has performed best with CRFs: in each round of active learning, $\mathcal{L}$ is sampled $|\mathcal{L}|$ times (with replacement) to create a unique modified labeled set $\mathcal{L}^{(c)}$. This is done C times to create C unique labeled sets. Then a committee of C models is trained: Each model $\theta^{(c)}$ is trained using its corresponding labeled set $\mathcal{L}^{(c)}$. Then the disagreement among the committee members about labeling an instance is measured as its informativeness:

$$\phi^{QBC}(\vec{x}) = -\sum_{\vec{y}\epsilon\mathcal{N}^C} P(\vec{y}|\vec{x};C)\, log\, P(\vec{y}|\vec{x};C). \quad (5)$$

In this equation, $\mathcal{N}^C$ is the union of N-best labelings of all models in the committee, and $P(\vec{y}|\vec{x};C) = \frac{1}{C}\sum_{c=1}^{C} P(\vec{y}|\vec{x};\theta)$ is the consensus posterior probability for some label sequence $\vec{y}$.

2.3 Representativeness

It is suggested that considering representativeness of instances can reduce the chance of selecting outliers in the process of active learning (Roy and McCallum, 2001). Representativeness can be measured by density of each instance, defined as the average similarity of an instance to other instances. Because the computation of density can be quite time-consuming in large-scale data sets, it is suggested to compute density in clusters (Tang et al., 2002; Shen et al., 2004) or in a k-Nearest-Neighbor manner (Zhu et al., 2008). Representativeness is applied by multiplying density to any arbitrary uncertainty measure to prevent outliers. Settles and Craven (2008) define a query strategy based on density:

$$\phi^{ID}(\vec{x}) = \phi^{LC}(\vec{x}) \times [ID(\vec{x})]^{\beta}, \quad (6)$$

$$ID(\vec{x}) = \frac{1}{U}\sum_{u=1}^{U} Sim(\vec{x}, \vec{x}^{(u)}). \quad (7)$$

Parameter β controls the relative effect of density $ID(\vec{x})$. This density uses a similarity measure $Sim(\cdot,\cdot)$ to compute the average similarity of an instance with all other unlabeled instances. The similarity measure used by Settles and Craven (2008) is a cosine similarity between two instances after being transformed to a vector of fixed length using this relation:

$$\vec{x} = \left[\sum_{t=1}^{T} f_1(x_t), \ldots, \sum_{t=1}^{T} f_J(x_t)\right], \quad (8)$$

where $f_j(x_t)$ is the value of feature f_j for token x_t, and J is the number of features in input representation. These features can be generated using CRF feature templates. Please refer to Settles and Craven (2008) for more details.

2.4 Fisher Information

We also evaluate the FIR (Fisher Information Ratio) method proposed by Settles and Craven (2008). Two vectors based on Fisher information are defined:

$$\mathcal{I}_{\vec{x}}(\theta) = \sum_{\hat{y}\epsilon\mathcal{N}} P(\hat{y}|\vec{x};\theta)\left[\left(\frac{\partial \log P(\hat{y}|\vec{x};\theta)}{\partial\theta_1}\right)^2 + \delta, \dots, \left(\frac{\partial \log P(\hat{y}|\vec{x};\theta)}{\partial\theta_K}\right)^2 + \delta\right], \quad (9)$$

$$\mathcal{I}_{\mathcal{U}}(\theta) = \frac{1}{|\mathcal{U}|}\sum_{u=1}^{|\mathcal{U}|}\mathcal{I}_{\vec{x}(u)}(\theta), \quad (10)$$

where $\mathcal{I}_{\vec{x}}(\theta)$ and $\mathcal{I}_{\mathcal{U}}(\theta)$ are the Fisher information matrices for sequence $\vec{x}$ and unlabeled pool $\mathcal{U}$ respectively. These matrices are estimated using their diagonal due to performance issues. Also K is the total number of CRF features, $\mathcal{N}$ is the set of N-best label sequences for input $\vec{x}$ and constant δ<<1 is added to prevent division by zero. Finally, FIR measures the informativeness of instances using:

$$\phi^{FIR}(\vec{x}) = -trace\left(\mathcal{I}_x(\theta)^{-1}\mathcal{I}_{\mathcal{U}}(\theta)\right). \quad (11)$$

2.5 Instability Sampling

(Zhu and Ma, 2012) suggest selecting instances which are most unstable. They propose two new methods to select most unstable instances based on recent active learning cycles: label-insensitive instability sampling (LIIS) and label-sensitive instability sampling (LSIS). Given an unlabeled instance $\vec{x}$ at i[th] learning cycle, its instability value in LIIS is estimated by:

$$\phi^{LIIS}(\vec{x}) = \phi_i^{SE}(\vec{x}) + \sum_{i-l<k\leq i}\left(\phi_k^{SE}(\vec{x}) - \phi_{k-1}^{SE}(\vec{x})\right), \quad (12)$$

where $\phi_i^{SE}(\vec{x})$ is $\phi^{SE}(\vec{x})$ at i[th] learning cycle and l is the number of cycles considered for instability estimation. Likewise, the instability value of $\vec{x}$ in LSIS is estimated by:

$$\phi^{LSIS}(\vec{x}) = \phi_i^{SE}(\vec{x}) + \sum_{i-l<k\leq i}\delta\left(\vec{y}^{(k)},\vec{y}^{(k-1)}\right)\left(\phi_k^{SE}(\vec{x}) - \phi_{k-1}^{SE}(\vec{x})\right), \quad (13)$$

where $\delta\left(\vec{y}^{(k)},\vec{y}^{(k-1)}\right)$ is 0 if the predicted label sequences $\vec{y}^{(k)}$ and $\vec{y}^{(k-1)}$ are the same and 1 otherwise. It's worthwhile to point that none of the instability sampling methods have been evaluated in the context of sequence labeling and they have only been evaluated in the context of classification.

3 The First Proposed Method: Weighted Gradient Uncertainty (WGU)

The new method to be introduced in this article is an improvement over the FIR method (subsection 2.4). According to evaluations by Settles and Craven (2008), the FIR method didn't perform well in practice despite its sound theory. In this section, first we investigate the essence of each component of $\mathcal{I}_{\vec{x}}(\theta)$:

$$\mathcal{I}_{\vec{x}}(\theta)_k = \sum_{\hat{y}\epsilon\mathcal{N}} P(\hat{y}|\vec{x};\theta)\left(\frac{\partial\ log\ P(\hat{y}|\vec{x};\theta)}{\partial\theta_k}\right)^2. \quad (14)$$

According to this relation, the k[th] component of Fisher vector $\mathcal{I}_{\vec{x}}(\theta)$ is the weighted sum of squared gradients of log-probabilities for the N best labelings for instance $\vec{x}$ in k[th] dimension of CRF features. It can be seen intuitively that each component of the Fisher vector increases when there is a kind of entropy between the N-best probabilities. That's because when for example the best label sequence has probability 1 then its gradient will be zero in all dimensions (complete fit) and hence all the components will be zero. On the other hand, if N best label sequences have equal probabilities, none of them will have a zero gradient since none is a complete fit and $\mathcal{I}_{\vec{x}}(\theta)$ will be maximized.

To show this fact more rigidly, assume the N best label sequences as $\mathcal{N} = \{\vec{y}^{(1)}, \vec{y}^{(2)}, \dots, \vec{y}^{(N)}\}$, and also for simplicity, define: $P_n = P\left(\vec{y}^{(n)}\middle|\vec{x};\theta\right)$. Then we will have:

$$logP_n = \sum_{j=1}^{T}\sum_{i=1}^{K}\theta_i f_i\left(y_{j-1}^{(n)}, y_j^{(n)}, \vec{x}, j\right) - logZ_{\vec{\theta}}(\vec{x}), \quad (15)$$

and so, its partial derivative in k[th] dimension will be (assuming $\mathcal{N}$ contains all possible label sequences):

$$\frac{\partial log P_n}{\partial \theta_k} = \overbrace{\sum_{j=1}^{T} f_k\left(y_{j-1}^{(n)}, y_j^{(n)}, \vec{x}, j\right)}^{F_n^{(k)}} - \frac{1}{Z_{\vec{\theta}}(\vec{x})} \frac{\partial Z_{\vec{\theta}}(\vec{x})}{\partial \theta_k} = F_n^{(k)} - \sum_{m=1}^{N} P_m F_m^{(k)}, \quad (16)$$

where $F_n^{(k)}$ is the result of applying feature function f_k (from CRF model) on nth best label sequence. Now using (16) we can rewrite (14) as:

$$\mathcal{I}_{\vec{x}}(\theta)_k = \sum_{n=1}^{N} P_n \left(F_n^{(k)} - \sum_{m=1}^{N} P_m F_m^{(k)} \right)^2. \quad (17)$$

To fully understand each component, we further factorized the above relation and proved it to be equal to (the proof is omitted here for brevity):

$$\mathcal{I}_{\vec{x}}(\theta)_k = \sum_{i=1}^{N} \sum_{j=i+1}^{N} P_i P_j \left(F_i^{(k)} - F_j^{(k)} \right)^2. \quad (18)$$

This relation explains the meaning of components of the Fisher vector completely. Each component is a summation over N best label sequences. The expression under summation consists of two parts: $P_i P_j$ and $\left(F_i^{(k)} - F_j^{(k)}\right)^2$. It can be shown using Lagrange multipliers that the first part is maximized (independently) when $P_i = \frac{1}{N}, \forall i$; which means this part is maximized when maximum entropy between N best probabilities occurs. The second part is the squared difference of kth feature function applied to two label sequences. So this part is maximized when the dissimilarities between every two label sequences in N-best list are maximum, which in turn means the model has maximum uncertainty in choosing the N-best label sequences for the input. Notice that in this interpretation we have assumed the two parts to be independent while they are not actually. However since the number of features of CRF (i.e. K) is too large, the dependency is negligible and can be ignored. So we conclude that each component of the Fisher vector $\mathcal{I}_{\vec{x}}(\theta)$ is a measure of uncertainty of the model about the sequence $\vec{x}$ in the corresponding dimension. Accordingly, each component of the total Fisher information vector $\mathcal{I}_{\mathcal{U}}(\theta)$ is the average uncertainty of the model in the corresponding dimension.

Knowing the precise identity of Fisher vector $\mathcal{I}_{\vec{x}}(\theta)$, we propose a natural measure which we call Weighted Gradient Uncertainty (WGU) based on the facts explained in the previous paragraph:

$$\phi^{WGU}(\vec{x}) = \sqrt{\sum_j \mathcal{I}_{\mathcal{U}}(\theta)_j \left(\mathcal{I}_{\vec{x}}(\theta)_j\right)^2}. \quad (19)$$

This measure is the weighted norm of $\mathcal{I}_{\vec{x}}(\theta)$ with the total Fisher information vector $\mathcal{I}_{\mathcal{U}}(\theta)$ as the weight vector. This query strategy favors instances with high uncertainty in each dimension of CRF feature space, especially the dimensions where the average uncertainty is higher. In other terms, the WGU measure maximizes the components of the Fisher vector, while the FIR method minimizes the inversed components of the Fisher vector; and since many components of the Fisher vector are zero or near-zero, their inversed values are very large and block out the other larger components (with very small inverse values) leading to a measure which effectively just counts the number of zero components and chooses the instance with the maximum number of zero components.

4 Using Local Density for Noisy Conditions

As described in Introduction, a great issue in SLU systems is the presence of noise in utterances. To address this problem, all the ATIS[1] instances were converted to vectors according to (8) and were reduced to 2 dimensions using Principle Component Analysis (PCA). Then the global density $ID(\vec{x})$ for

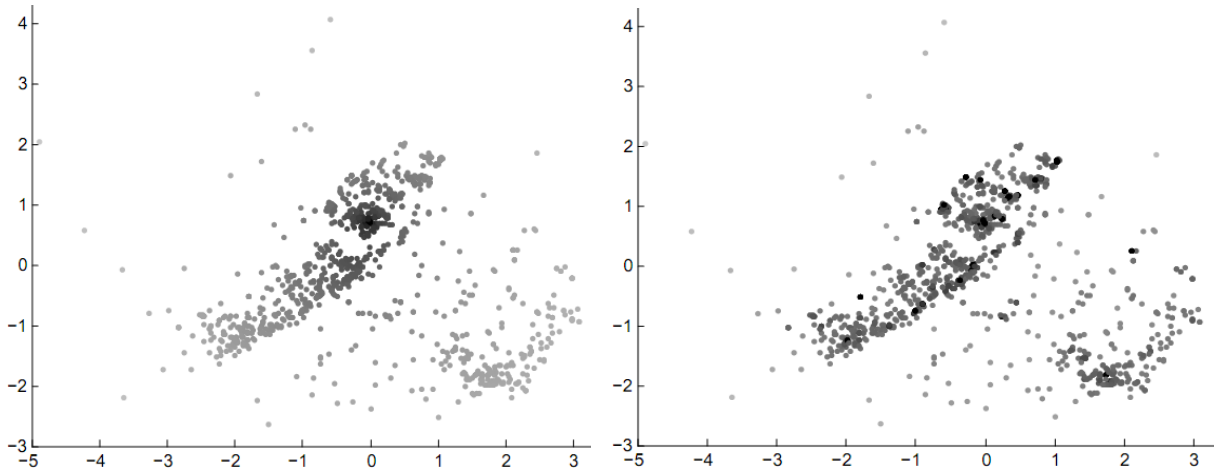
[1] ATIS is the dataset used in this article for evaluation; please read subsection 5.1.

each instance was computed using (7). Figure 2 shows the plot of all instances with darker points indicating instances with higher densities and lighter points showing the ones with lower densities.

As seen in Figure 2(a), the center of the distribution in terms of density is the darkest part. Also, the distribution of instances is not uniform at all, and excluding any part of the distribution especially parts further from the density center can lead to great decrease in performance of the model. The query strategy ϕ^{ID} (6) uses this density to reduce the chance of querying outliers. However, outliers as well as many other instances which are far from the density center are almost deprived of the chance of being selected. To address this problem and yet avoid outliers we choose to compute information density for each instance locally, i.e. using k nearest instances and not all instances. Thus, we define the local information density measure as follows:

$$LD(\vec{x}, k) = \frac{1}{k} \sum_{\vec{x}' \in \Gamma_k(\vec{x})} \text{Sim}(\vec{x}, \vec{x}'). \tag{20}$$

In which, $\Gamma_k(\vec{x})$ is the set of k most similar instances to $\vec{x}$, and k is the degree of locality.



(a) (b)

Figure 2. Plot of all ATIS instances. Darker points show higher densities and lighter ones show lower densities. (a) Using global density measure (b) Using local density measure (k=5).

The same procedure to plot Figure 2(a) is repeated again but with $LD(\vec{x}, k)$ computed as the density of each instance and the result is shown in Figure 2(b). The degree of locality is set to k=5. As seen in this plot, outliers are still completely grey which means they are avoided. Also, any small neighborhood with sufficient density is biased to black, which means the instances in the center of that neighborhood have almost the same chance of being queried as the instances in the center of global density (ID) in Figure 2(a).

Another advantage of local density is that it avoids noisy instances. Noisy instances in the SLU context are the utterances in which one or more words are erroneous due to ASR or user errors. Because of such errors, noisy instances take a small distance from their similar instances and reside alone in small neighborhoods.

Based on the LD measure (20), two active learning methods are considered: the first method applies local density measure to query strategy by multiplication (same as ϕ^{ID}):

$$\phi^{LD}(\vec{x}) = \phi^{LC}(\vec{x}) \times [LD(\vec{x}, k)]^{\beta}. \tag{21}$$

The second method which is proposed in this paper, strictly applies the local density measure by first filtering out instances with local densities lower than a threshold T, and then queries the most informative instance according to a certain query strategy (here we use ϕ^{LC}). This method is called Strict LD (SLD). We believe that this method of utilizing density measures is more effective than the traditional method (i.e. multiplying density measure by uncertainty measure (6)), since it does not affect all instances but

only very low-density ones. The threshold T is assumed to be in the form of $\alpha * \overline{LD}$, where $\overline{LD}$ is the average of local density over all unlabeled instances, and parameter α sets the intensity of filtering.

It is necessary to note that the k-Nearest-Neighbor density measure (Zhu et al., 2008) is identical to local density in definition but the motivation is different and in this article we look at the k-nearest-neighbor density from a completely different perspective: to avoid a shortcoming in the global density which is ignoring great parts of the input distribution and also to detect noisy instances.

5 Experiments

Experiments are all performed on the ATIS[1] data set (Hemphil et. al, 1990), both in noise-free and noisy conditions. In this section, the noise model used to generate noise is briefly described and then the evaluations are presented.

5.1 ATIS and Noise Model

ATIS is a relatively simple corpus which contains air travel information data. This corpus is the most commonly used data set for SLU research (Tur et. al, 2010). The data set contains questions (utterances) about flight, airport, and airline information. We specifically use the class-A (context independent) utterances from ATIS-3 corpus (Dahl et. al, 2004). These utterances are not semantically labelled, instead for each utterance there is an SQL command which queries the answer to the utterance from database. Thus a flat sematic representation was designed and semantic label sequences were generated semi-automatically from the SQL queries (as explained by He and Young (2006)). The flat semantic representation is listed in Table 1(a). A flat semantic representation is in fact a set of attributes (semantic labels) which are used to label an input utterance. Table 1(c) shows a typical utterance with sematic labels; note that IOB labeling scheme is used. Totally there are 1630 class-A instances (test + train) in ATIS-3 which are used in the experiments.

Attribute	Description	Attribute	Description
DCity	depart. city	ACity	arrival city
SCity	stop city	DAir	depart. airport
DDate	depart. date	ADate	arrival date
RDate	return date	AAir	arrival airport

(a)

Origin	Pair
ASR	via → fly at
Human	to Chicago → chica to Chicago
ASR	phoenix → t x

(b)

Show	flights	from	Denver	to	Washington	on	Sunday	arriving	before	noon
O	O	O	DCity	O	ACity	O	DDate	O	ADate-I	ADate-I

(c)

Table 1: (a) The flat semantic representation used to label utterances in the data set. (b) Some example pairs in noise model. Each pair is extracted from actual errors in ATIS-3 utterances. (c) A typical example from ATIS utterances.

Utterances in ATIS are de-noised by wizards[2]. There are two origins of noise: human (end-user) errors and ASR recognition errors. We design a simple noise-model based on actual errors and regenerate human and ASR errors. In ATIS-3, human errors are marked in SRO files and ASR errors are in N-best lists in log files. The noise model is a list of pairs of the form [correct-expression] → [erroneous-expression] which are applied to ATIS instances to add arbitrary percentage of noise. A few example pairs in the noise model are listed in Table 1(b). Each pair is extracted from an actual error; for example [phoenix] → [t x] is a result of an ASR error in ATIS-3 logs where "phoenix" in "Show me flights from phoenix …" was recognized as "t x" mistakenly. Obviously this pair is only applicable to an utterance which contains the word "phoenix".

[1] Air Travel Information System

[2] A wizard is a human expert who transcribes utterances or answers them (Hemphil et. al, 1990).

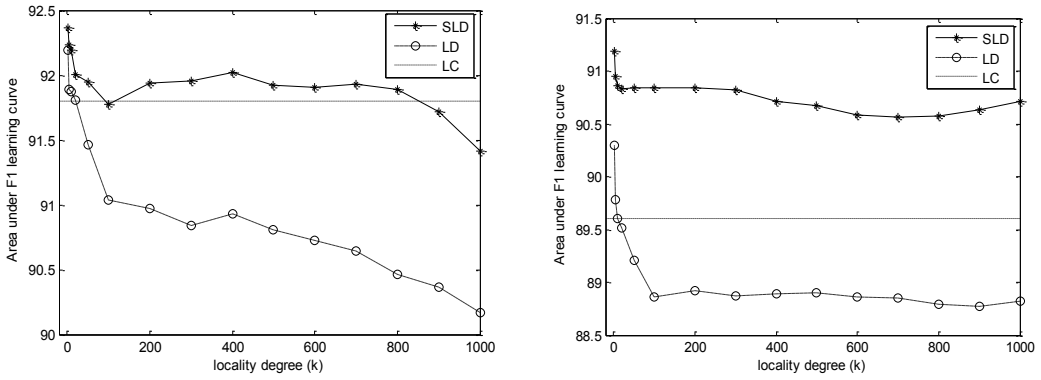
5.2 Parameter Settings

Using the noise model described, 3 levels of noise were generated: 7% of instances in level 1, 15% in level 2, and 25% in level 3 are noisy. In noisy conditions, when a noisy instance is selected by an active learning method, we assume that the instance is correctly detected as noisy by the annotator and is rejected (i.e. not added to $\mathcal{L}$); but the determination of an instance as noisy incurs a cost which we assume to be a quarter of cost of labeling one instance[1]. In all experiments, $\mathcal{L}$ is initialized with 5 random training instances. Batch size in all experiments is set to B=2 and new instances are added to $\mathcal{L}$ until the total labeling cost reaches 100. For query-by-committee method, we set C=4 and N=20 to balance between speed and accuracy. For LD and SLD, we set k=1 because it achieved best performance. For LIIS and LSIS, we set l=2 which achieved better results. Each method is evaluated as the average of 5 trials and each trial is performed using 5-fold cross validation. The reported performance for each method is the area under F1 learning curve (F1 score in SLU is computed as described by Tur and De Mori (2011)).

5.3 Effect of Locality

By initial evaluations, β=1 and α=0.6 were chosen for the LD and SLD method respectively. In Figure 3, the performances of LD and SLD for different degrees of locality (for k=1 to 1000) are shown. The performance of the LC method is also shown for comparison.

As seen in Figure 3, local density improves uncertainty measure (i.e. ϕ^{LC}, which is the base method in LD and SLD) and performs better than global density (i.e. local density with k=1000+). Note that LD has led to better performances than LC only for very local densities (i.e. k<5) while SLD has improved the performance of LC almost for all degrees of locality. It can also be seen that applying density strictly is more effective than the traditional way for all degrees of locality especially in noisy conditions.



(a) Noise-free condition (b) Noisy condition (average of all levels)

Figure 3. Effect of locality degree (in computation of information density) on performance of active learning methods. Plots (a) and (b) show the area under F1 learning curve for different values of k in LD and SLD methods, for noise-free and noisy conditions respectively. The area under F1 learning curve for LC is also shown for comparison.

5.4 Evaluations

The detailed results of the discussed active learning methods on different levels of noise are presented in Table 2. In each row, best performance is bolded and underlined, and second best performance is just bolded. Random refers to the random sampling of instances (passive learning). In noise-free condition, LD and SLD have improved a little over LC, but in average, SLD has performed remarkably better than LC, which shows the effectiveness of using local density to avoid noisy instances (note that LC is the base method used in LD and SLD). The instability sampling methods have improved over uncertainty

[1] The cost of labelling one instance is equal to 1 for any instance. In this paper, learning curves are depicted in terms of annotation cost which is equivalent to annotation time (please refer to Tomanek and Hahn (2010)).

sampling (i.e. SE) but not significantly. In the last row of Table 2 the running time of one cycle of active learning for each method is presented in seconds. QBC is the slowest method and LC is the fastest one. WGU is the second best in average performance but is rather slow in comparison to LC and this is a disadvantage of WGU. In fact all methods that iterate over best labelings are considerably slower that LC.

Learning curves cannot be shown for all active learning methods due to lack of space. Instead, learning curves are shown for selected methods. In Figure 4, learning curves for five methods of SLD, WGU, LIIS, FIR, and random are shown. It can be seen that the new WGU method has the best performance in early stages of active learning but soon declines and stays above the curve of LIIS. Also, the difference of SLD with other methods is more remarkable in the noisy conditions.

	Random	LC	SE	QBC	ID	FIR	LIIS	LSIS	WGU	LD	SLD
Noise-free	84.5	91.8	91.9	90.5	90.2	89.5	91.7	91.8	**92.1**	**92.1**	**92.4**
Noise level 1	84.1	**91.5**	90.7	90	89.6	89.1	91.1	90.8	**91.7**	91.2	**91.7**
Noise level 2	83.2	88.9	89	88.2	88.1	89.2	89.4	88.9	**90.4**	89.4	**91.1**
Noise level 3	83	88.4	88.4	87.5	87.7	**88.9**	88.7	87.8	**90**	89.3	**91**
Average	83.7	90.1	90	89	88.9	89.2	90.2	89.8	**91.1**	90.5	**91.6**
Runtime	**5**	**5**	8	20	**5.5**	8	8	8	8	**5.5**	**5.5**

Table 2. Area under F1 learning curves (max possible score is 100) and runtimes of various active learning methods on different levels of noise.

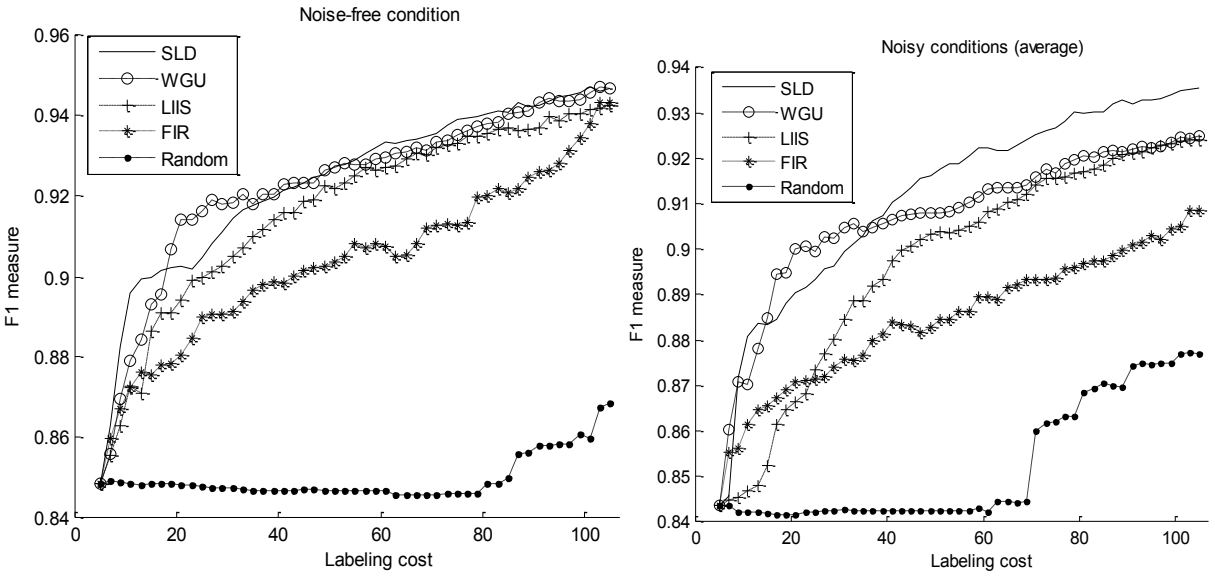


Figure 4. Learning curves for five selected methods: SLD, WGU, LIIS, FIR, and random for noise-free and noisy conditions (averaged across noise levels 1-3). Each learning curve shows the F1 measures achieved by the corresponding method for different labelling costs up to 100.

6 Conclusion

In this paper, best known active learning methods applicable to sequence labeling tasks were evaluated in the field of SLU (Spoken Language Understanding) in real conditions of noise. The new method of WGU (Weighted Gradient Uncertainty) with theoretical justification was proposed and performed well in the evaluations. Also, to deal directly with noisy instances, two methods of LD (Local Density) and SLD (Strict LD) were proposed based on the local density concept. It is possible to apply local density to WGU or other methods to achieve even better results but this could be the subject of future work.

References

Burr Settles and Mark Craven. 2008. *An Analysis of Active Learning Strategies for Sequence Labeling Tasks*, In EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1070-1079.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. *The ATIS spoken language systems pilot corpus*. In Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 96-101.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. *Expanding the scope of the ATIS task: the ATIS-3 corpus*. In Proceedings of the workshop on Human Language Technology (HLT '94). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43-48.

Gokhan Tur, Dilek Hakkani-Tür, and Larry P. Heck. 2010. *What is left to be understood in ATIS?* IEEE Spoken Language Technology Workshop (SLT), Berkeley, California, USA, December 12-15, pp. 19-24.

Gokhan Tur and Renato De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, First Edition, John Wiley & Sons.

Gokhan Tur, Marzin Rahim, and Dilek Hakkani-Tür. 2003. *Active Learning for Spoken Language Understanding*, In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 276-279.

Isabelle Jars and Franck Panaget. 2008. *Improving Spoken Language Understanding with information retrieval and active learning methods*, In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5001-5004.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, In Proceedings of the International Conference on Machine Learning (ICML), pp. 282–289.

Jingbo Zhu and Matthew Ma. 2012. *Uncertainty-based active learning with instability estimation for text classification*. ACM Transactions on Speech and Language Processing (TSLP), vol. 8(4) - 01/2012.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. 2008. *Active learning with sampling by uncertainty and density for word sense disambiguation and text classification*. In Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08), Vol. 1, pp. 1137-1144.

Katrin Tomanek and Udo Hahn. 2010. *A comparison of models for cost-sensitive active learning*. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 1247-1255.

Lynette Hirschman. 1992. *Multi-Site Data Collection for a Spoken Language Corpus,* In Proceedings of International Conference on Spoken Language Processing, Banff, Canada.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. *Active learning for statistical natural language parsing*, In Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting, pp.120–127.

Nicholas Roy and Andrew McCallum. 2001. *Toward optimal active learning through sampling estimation of error reduction*, In Proceedings of the International Conference on Machine Learning (ICML), pp. 441–448.

Pierre Gotab, Frédéric Béchet, and Géraldine Damnati. 2009. *Active learning for rule-based and corpus-based Spoken Language Understanding models*, In Proceedings of IEEE Conference on Automatic Speech Recognition and Understanding, pp. 444-449.

Yulan He and Steve Young. 2005. *Semantic processing using the hidden vector state model*, Computer Speech & Language, vol. 19, no. 1, pp. 85–106.