Confidence-Based Stopping Criteria for Active Learning for Data Annotation

JINGBO ZHU and HUIZHEN WANG
Northeastern University
EDUARD HOVY
University of Southern California
and
MATTHEW MA
Scientific Works

The labor-intensive task of labeling data is a serious bottleneck for many supervised learning approaches for natural language processing applications. Active learning aims to reduce the human labeling cost for supervised learning methods. Determining when to stop the active learning process is a very important practical issue in real-world applications. This article addresses the stopping criterion issue of active learning, and presents four simple stopping criteria based on confidence estimation over the unlabeled data pool, including *maximum uncertainty*, *overall uncertainty*, *selected accuracy,* and *minimum expected error* methods. Further, to obtain a proper threshold for a stopping criterion in a specific task, this article presents a strategy by considering the label change factor to dynamically update the predefined threshold of a stopping criterion during the active learning process. To empirically analyze the effectiveness of each stopping criterion for active learning, we design several comparison experiments on seven real-world datasets for three representative natural language processing applications such as word sense disambiguation, text classification and opinion analysis.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms, Experimentation, Theory, Language

Additional Key Words and Phrases: Active learning, uncertainty sampling, stopping criterion, confidence estimation, text classification, word sense disambiguation

This research was supported by the National Science Foundation of China (60873091).
Authors' addresses: J. Zhu and H. Wang, Key Laboratory of Medical Image Computing (Ministry of Education), Northeastern University, Shenyang, Liaoning, P. R. China, 110004; email: {zhujingbo,wanghuizhen}@mail.neu.edu.cn; E. Hovy, USC Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292-6695; email: hovy@isi.edu; M. Ma, Scientific Works, Tiffany Court, Princeton Junction, NJ 08550; email: mattma@ieee.org.

DOI 10.1145/1753783.1753784 http://doi.acm.org/10.1145/1753783.1753784

ACM Reference Format:
Zhu, J., Wang, H., Hovy, E., and Ma, M. 2010. Confidence-based stopping criteria for active learning for data annotation. ACM Trans. Speech Lang. Process. 6, 3, Article 3 (April 2010), 24 pages.
DOI = 10.1145/1753783.1753784 http:/doi.acm.org/10.1145/1753783.1753784

1. INTRODUCTION

In machine learning approaches to Natural Language Processing (NLP), supervised learning methods generally set their parameters using labeled training data. However, creating a large labeled corpus is often expensive and time consuming in some real-world applications, and is a bottleneck to build an effective supervised classifier for a new application or domain. For example, consider the word sense disambiguation task. In this case, building a large sense-tagged corpus is critical for good performance but requires significant efforts of human annotators, as described in the OntoNotes [Hovy et al. 2006].

The goal of *active learning* is to design a learning algorithm which has the ability to automatically select the most informative unlabeled examples for human annotation [Cohn et al. 1994; Seung et al. 1992]. The ability of the active learner is also referred to as selective sampling. Active learning aims to minimize the amount of human labeling effort required for a supervised classifier to achieve a satisfactory performance [Cohn et al. 1996]. In recent years active learning has been widely studied in various Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (WSD) [Chen et al. 2006; Zhu and Hovy 2007; Chan and Ng 2007], Text Classification (TC) [Lewis and Gale 1994; McCallum and Nigam 1998a; Tong and Koller 2001], Named Entity Recognition (NER) [Shen et al. 2004; Jones 2005; Tomanek et al. 2007], chunking [Ngai and Yarowsky 2000], Information Extraction (IE) [Thompson et al. 1999; Culotta and McCallum 2005], and statistical parsing [Hwa 2000; Tang et al. 2002; Becker and Osborne 2005].

We focus on *pool-based active learning* [Lewis and Gale 1994] in which the learner chooses the most informative unlabeled instances from a pool of unlabeled instances for human labeling. Two other popular variants widely used in previous active learning studies include *stream-based active learning* [Freund et al. 1997] and *active learning with membership queries* [Angluin 1988]. In the pool-based active learning setting, the learner is presented with a fixed pool of unlabeled instances, whereas in the stream-based active learning setting, the learner is presented with a stream of unlabeled instances. From this perspective, stream-based active learning can be viewed as an online version of the pool-based model. The active learning with membership queries model can be viewed as a pool-based case where the pool consists of all possible points in a domain [Baram et al. 2004]. We do not consider these variations in this work.

In the pool-based active learning setting, two major schemes exist: *uncertainty sampling* and *committee-based sampling*. Uncertainty sampling [Lewis and Gale 1994] uses only one classifier to identify unlabeled examples on which the classifier is least confident. Committee-based sampling [Seung et al. 1992; Dagan and Engelson 1995; McCallum and Nigam 1998a] generates a committee of classifiers and selects the next unlabeled example by the principle of

maximal disagreement among these classifiers, which is uncertainty sampling with an ensemble. By using these selective sampling techniques, the size of the labeled training data can be significantly reduced for text classification [Lewis and Gale 1994; McCallum and Nigam 1998a] and word sense disambiguation [Chen et al. 2006; Zhu and Hovy 2007].

In active learning applications, obtaining of a stopping criterion is a very important practical issue because it makes little sense to continue the active learning procedure until the entire unlabeled corpus has been labeled. In principle, how to learn a stopping criterion is a problem of estimation of classifier effectiveness during active learning [Lewis and Gale 1994]. That is, defining an appropriate stopping criterion for active learning is a trade-off issue between labeling cost and effectiveness of the classifier.

This article presents four simple confidence-based criteria for active learning, including *maximum uncertainty, overall uncertainty, selected accuracy,* and *minimum expected error* methods. To obtain a proper threshold for each stopping criterion in a specific active learning task, this article further presents a threshold update strategy that uses the label change factor to dynamically update the predefined threshold of a stopping criterion. The label change factor will be introduced in Section 4.5. These proposed methods are easy to implement and involve only small additional computation costs. They can also be easily applied to several different learners, such as Naïve Bayes (NB) and Maximum Entropy (MaxEnt). Finally, we evaluate the effectiveness of these stopping criteria for active learning on three NLP tasks including word sense disambiguation, text classification, and opinion analysis, using seven real-world datasets.

2. ACTIVE LEARNING PROCESS

In the pool-based active learning framework, a small number of labeled samples and a large number of unlabeled examples are first collected in the initialization stage, and a closed-loop stage of query (i.e., selective sampling process) and retraining is adopted. In this article, we are interested in *uncertainty sampling* schemes [Lewis and Gale 1994] for pool-based active learning, which in recent years has been widely studied in tasks such as word sense disambiguation [Chen et al. 2006; Chan and Ng 2007], Text Classification (TC) [Lewis and Gale 1994; Zhu et al. 2008b], statistical syntactic parsing [Tang et al. 2002], and named entity recognition [Shen et al. 2004].

In uncertainty sampling schemes, an unlabeled example x with maximum uncertainty is chosen for human annotation at each learning cycle. The maximum uncertainty implies that the current classifier (i.e., the learner) has the least confidence on its classification of this unlabeled example. The main difference among the various pool-based active learning algorithms is the method of assessing the uncertainty of each unlabeled example in the pool. In the case of probabilistic models, the uncertainty of the classifier is commonly estimated using the entropy of its output [Tang et al. 2002; Chen et al. 2006; Zhu and Hovy 2007]. For active learning with nonprobabilistic models such as support vector machines [Tong and Koller 2001; Schohn and Cohn 2000], the classification

Procedure: Active Learning Process

Input: Initial small training set L, and pool of unlabeled data set U

Use L to train the initial classifier C

Repeat

1. Use the current classifier C to label all unlabeled examples in U
2. Based on uncertainty sampling scheme, select m[1] most uncertain unlabeled examples from U, and ask an oracle for labeling
3. Augment L with these m new labeled examples, and remove them from U
4. Use L to retrain the current classifier C

Until the predefined stopping criterion SC is met.

Fig. 1. Active learning with uncertainty sampling.

margin is used. In this article, we mainly focus on the problem of applying a stopping criterion for active learning with probabilistic models, but will also discuss the possibility of applying to nonprobabilistic models, to be described in Section 4.6.

In uncertainty sampling schemes, the uncertainty measurement function based on the entropy is expressed by [Tang et al. 2002; Chen et al. 2006; Zhu and Hovy 2007]

$$UM(x) = -\sum_{y \in Y} P(y \mid x) \log P(y \mid x), \tag{1}$$

where $P(y \mid x)$ is the a posteriori probability. We denote the output class $y \in Y = \{y_1, y_2, \ldots, y_k\}$. $UM(.)$ denotes the uncertainty measurement function based on the entropy estimation of the classifier's posterior distribution. A higher $UM(x)$ value indicates that the unlabeled example x is more uncertain from the viewpoints of the classifier.

3. PROBLEMS OF GENERAL STOPPING CRITERION

As shown in Figure 1, the active learning process repeatedly provides the most uncertain unlabeled examples to an oracle for labeling, and updates the training set, until the predefined stopping criterion SC is met. The goal for using active learning is to expedite the learning process and reduce the manual labeling efforts. In this case, we can define a stopping criterion by means of determining when the classifier has reached the maximum effectiveness during the active learning procedure.

Along this line of thinking, a general stopping criterion SC can be defined by

$$SC_{AL} = \begin{cases} 1 & effectiveness(C) \geq \theta \\ 0 & otherwise, \end{cases} \tag{2}$$

where θ is a user predefined constant and the function $effectiveness(C)$ evaluates the effectiveness of the current classifier. The learning process ends only

[1]A batch-based sample selection labels the top-m most uncertain unlabeled examples at each learning cycle to decrease the number times the learner is retrained.

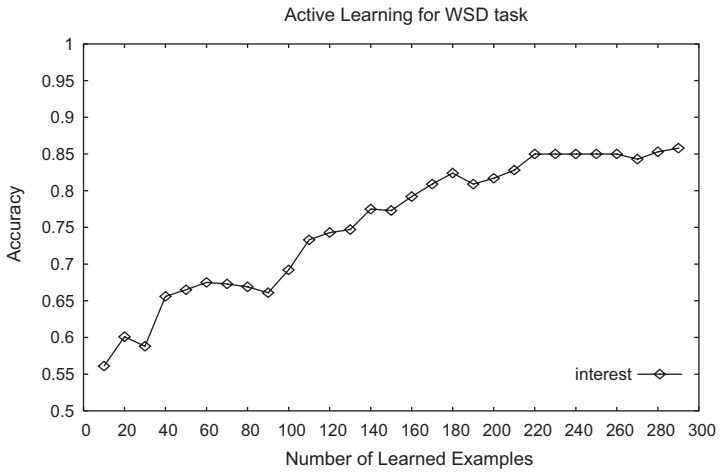

Fig. 2. An example of active learning for WSD on word "*interest*".

if the stopping criterion function SC_{AL} is equal to 1. The value of constant θ represents a trade-off between the cost of annotation and the effectiveness of the resulting classifier. A larger θ would result in more unlabeled examples to be selected for human annotation, and the resulting classifier would be more robust. A smaller θ, on the other hand, means fewer unlabeled examples being selected to annotate and the resulting classifier will be less robust.

There are three common ways to define the function *effectiveness*(C) [Li and Sethi 2006] as follows.

—The active learning process can end if the labeled training set reaches desirable size. However, it is almost impossible to predefine an appropriate size of desirable labeled training data that is guaranteed to induce the most effective classifier.

—The active learning loop can end if no uncertain unlabeled examples can be found in the pool. That is, all informative examples have been selected for human labeling from the viewpoints of the classifier. This situation seldom occurs in practice.

—The active learning process can end if the targeted performance level is achieved. However, it is difficult to predefine an appropriate and achievable performance, since this should depend on the problem at hand and the users' requirements.

To overcome these difficulties, it seems an appealing solution to stop the active learning process when repeated cycles show no significant performance improvement during active learning. That is, the classifier's performance change would be a good signal to define a stopping criterion for active learning. Here we give an example to explore using the performance change factor to define a proper stopping criterion, as shown in Figure 2.

As shown in Figure 2, the accuracy performance generally increases, but apparently degrades at iterations 30, 90, and 190, and stablizes during iterations

220–260 during the active learning process. In this example, the actual highest performance of 91.5% is achieved after 900 annotated examples, which is not shown. The accuracy performance curve shows an increasing trend, but it is not monotonically increasing. In such a case, it will be not easy to automatically determine the point where no significant performance improvement can be further achieved by simply looking at the trend of performance change. A false judgement can often be misled by the nonmonotonic behavior of the performance curve.

To sufficiently estimate the performance of the classifier during active learning, a separate validation set should be prepared in advance. However, in the separate validation set, too few samples may not be adequate for a reasonable estimation and may result in an incorrect result, and too many samples would cause additional high cost because the separate validation set is generally constructed manually in advance.

In a real-world active learning application, preparing an appropriate separate validation set for performance estiamtion is often infeasible due to too high manual annotation cost involved. In addition, cross-validation on the labeled set is also almost impractical during the active learning procedure, because the alternative of requiring a held-out validation set for active learning is counterproductive. To remedy these problems, in this study we look for a self-contained method to define a proper stopping criterion for active learning.

4. CONFIDENCE-BASED STOPPING CRITERIA

To define an appropriate stopping criterion for active learning, we consider effectiveness estimation as the second task to confidence estimation of the current classifier. If the classifier already has sufficient confidence on its classification of the remaining unlabeled data, we can assume the current labeled data is sufficient to train the classifier with maximum effectiveness. In other words, attempting to obtain the labels of these remaining unlabeled examples is not going to significantly improve learner performance.

Based upon such assupmption, this article presents four simple confidence-based stopping criteria for pool-based active learning, including *maximum uncertainty, overall uncertainty, selected accuracy,* and *minimum expected error* methods. Parts of this work were originally introduced in our previous studies [Zhu and Hovy 2007; Zhu et al. 2008a, 2008b]. To obtain a proper threshold for a confidence-based stopping criterion in a specific task, we further describe a threshold update strategy which uses the classification change factor to dynamically update the predefined threshold of a stopping criterion during the active learning process.

4.1 Maximum Uncertainty Method

In uncertainty sampling schemes, the most uncertain unlabeled example is viewed as the most informative instance to be chosen by the learner at each learning cycle [Lewis and Gale 1994]. The uncertainty value of the chosen example is a good signal to reflect the confidence of the current classifier on all unlabeled examples. If the uncertainty value of this chosen example is

sufficiently small, we can assume that the current classifier has sufficient confidence on its classification of the remaining unlabeled data. Therefore, the active learning process can stop.

Based on this assumption, we present an approach based on uncertainty estimation of the most uncertain unlabeled example to obtain a stopping criterion for active learning, named the *Maximum Uncertainty* (MU) method. Its strategy is to consider whether the uncertainty values of all unlabeled example are less than a very small predefined threshold.[2] The stopping criterion SC_{MU} can be defined by

$$SC_{MU} = \begin{cases} 1 & \forall x \in U, UM(x) \leq \theta_{MU} \\ 0 & otherwise, \end{cases} \tag{3}$$

where θ_{MU} is a user-predefined uncertainty threshold, and U denotes the unlabeled pool.

4.2 Overall Uncertainty Method

The motivation behind the *Overall Uncertainty* (OU) criterion is similar to that of the maximum uncertainty criterion. However, the maximum uncertainty criterion only considers the most uncertain example at each learning cycle. The overall uncertainty method considers the overall uncertainty on all unlabeled examples. If the overall uncertainty of all unlabeled examples becomes very small, we can assume that the current classifier has sufficient confidence on its classification of the remaining unlabeled data. The strategy of the OU method is to consider whether the average uncertainty value of all remaining unlabeled examples is less than a very small predefined threshold. The stopping criterion SC_{OU} can be defined by

$$SC_{OU} = \begin{cases} 1 & \frac{\sum_{\forall x \in U} UM(x)}{|U|} \leq \theta_{OU} \\ 0 & otherwise, \end{cases} \tag{4}$$

where θ_{OU} is a user-predefined uncertainty threshold, and $|U|$ denotes the size of the unlabeled data pool U.

4.3 Selected Accuracy Method

In batch mode active learning settings, the classification accuracy on the top-m selected examples (i.e., m most uncertain cases) at each learning cycle would be a good signal to indicate the confidence of the current classifier on remaining unlabeled examples. It is therefore easy to estimate this accuracy based on the feedback from the oracle when an active learner asks for true labels for these selected unlabeled examples (see Figure 1). The current classifier shall have the least confidence on its classifications of these chosen unlabeled examples at each learning cycle. If the current classifier can correctly classify these chosen

[2]To determine an appropriate threshold for a stopping criterion, some experiments were conducted to demonstrate the effectiveness of the stopping criterion with various thresholds, discussed in Section 5.

unlabeled examples, we can assume that the classifier has sufficient confidence on its classification of the remaining unlabeled data.

Based on this assumption, we present a *Selected Accuracy* (SA) method that is based on the feedback from the oracle. Its strategy is to consider whether the current classifier can correctly predict the labels of top-m selected unlabeled examples. In other words, if the accuracy performance of the current classifier on these most uncertain examples is larger than a predefined threshold, the active learning process can stop. The stopping criterion SC_{SA} can thus be defined by

$$SC_{SA} = \begin{cases} 1 & ACC_m(C) \geq \theta_{SA} \\ 0 & otherwise, \end{cases} \tag{5}$$

where θ_{SA} is a user-predefined accuracy threshold and function $ACC_m(C)$ evaluates the accuracy performance on the top-m selected unlabeled examples through feedback of the Oracle.

4.4 Minimum Expected Error Method

So far MU, OU, and SA methods do not directly reflect the effectiveness of the classifier, as they only consider either the uncertainty of each unlabeled example or the accuracy of the top-m selected unlabeled examples at each iterative step. In previous work, Roy and McCallum [2001] presented a method to select the most informative unlabeled example by optimizing expected future error on future unlabeled examples during an active learning process. It is believed that the expected error of a classifier is closely related to the effectiveness of the classifier. In this article we present a statistical learning approach to defining a stopping criterion which is based on the estimation of the current classifier's expected error on all future unlabeled examples, named the *Minimum Expected Error* (MEE) method. The motivation behind MEE is that a classifier C with maximum effectiveness results in the lowest expected error on whole test set in the learning process. The stopping criterion SC_{MEE} can be defined as

$$SC_{MEE} = \begin{cases} 1 & Error(C) \leq \theta_E \\ 0 & otherwise, \end{cases} \tag{6}$$

where $Error(C)$ evaluates the expected error of classifier C that closely reflects the classifier effectiveness. θ_E is a user-predefined error threshold.

The key issue in defining the MEE-based stopping criterion SC_{MEE} is how to calculate the expected error of classifier C at each learning cycle. Given a labeled training set L and an input sample x, we can express the expected error of the classifier C by

$$Error(C) = \int R(C(x) \mid x)P(x)dx, \tag{7}$$

where $P(x)$ represents the known marginal distribution of x. $C(x)$ represents the classifier's decision that is one of k classes: $y \in Y = \{y_1, y_2, \ldots, y_k\}$. $R(y_i \mid x)$

denotes a conditional loss for classifying the input sample x into a class y_i that can be defined by

$$R(y_i \mid x) = \sum_{j=1}^{k} \lambda[i, j] P(y_j \mid x), \tag{8}$$

where $P(y_j \mid x)$ is the a posteriori probability produced by classifier C. $\lambda[i, j]$ is a zero-one loss function for every class pair $\{i, j\}$ that assigns no loss to a correct classification, and assigns a unit loss to any error.

In this study, we focus on pool-based active learning in which a large unlabeled data pool U is available, as described Figure 1. During the active learning process, our goal is to estimate a classifier's expected error on future unlabeled examples in the pool U in order to determine the active learning stopping criterion. The pool U can provide an estimate of $P(x)$. So for minimum error rate classification [Duda and Hart 1973] on unlabeled examples, the expected error of the classifier C can be rewritten as

$$Error(C) = \frac{1}{|U|} \sum_{x \in U} \left(1 - \max_{y \in Y} P(y \mid x)\right). \tag{9}$$

Assuming N unlabeled examples in the pool U, the total time is $O(N)$ for automatically determining whether the proposed stopping criterion SC_{MEE} is satisfied in the active learning. If the pool U is very large (e.g., more than 100,000 examples), it would result in high computation cost at each iteration of active learning.

A good approximation is to estimate the expected error of the classifier using a randomly chosen subset of the pool, instead of all unlabeled examples in U. Empirically, a good estimation of expected error can be formed with a few thousand examples [Roy and McCallum 2001].

4.5 Threshold Update Strategy

Perhaps there are different appropriate thresholds for a confidence-based stopping criterion in different active learning applications. Therefore, it is a challenge to predefine an appropriate threshold for each confidence-based stopping criterion in a specific task. To solve this problem, we present a *threshold update strategy* by considering a label change factor to dynamically update the predefined threshold during the active learning process.

Whereas the four confidence-based criteria described earlier directly reflect the confidence of the current classifier on all remaining unlabeled examples, we further explore the motivation behind uncertainty sampling, which is to find some unlabeled examples near decision boundaries, and use them to clarify the position of decision boundaries. In other words, in uncertainty sampling schemes, an unlabeled example with maximum uncertainty has the highest chance to change the decision boundaries.

If there is no unlabeled example that can potentially change the decision boundaries, getting the labels of these remaining unlabeled examples is unlikely to help the learner much, hence the active learning process should stop. The difficulty lies in how to exactly find which unlabeled example can truly

change the decision boundaries in the next learning cycle, because the true label of each unlabeled example is unknown.

To overcome this difficulty, we make a simple assumption that labeling an unlabeled example may shift the decision boundaries if this example was previously at "left" of the boundary and is now at "right", or vice versa. In other words, if an unlabeled example is assigned to two different labels during two adjacent learning cycles,[3] we assume that the labeling of this unlabeled example has a good chance to change the decision boundaries. Once there is no such unlabeled example in the remaining pool, we think the active learning process becomes stable, and can end.

Based on such assumption, we present a method, called the *Threshold Update (TU) strategy*, to automatically adjust the predefined threshold of a stopping criterion during the active learning process. This method considers the potential ability of each unlabeled example on changing decision boundaries and checks whether there is any classification label change to the remaining unlabeled examples during two recent consecutive learning cycles (previous and current). It checks whether the active learning becomes stable when the current stopping criterion is satisfied. If not, we believe there are some remaining unlabeled examples that can potentially shift the decision boundaries. In such cases, the threshold of the current stopping criterion can be revised to keep the active learning process going.

For the previously described four confidence-based stopping criteria such as MU, OU, SA, and MEE methods, the four corresponding strategies are given as follows.

—TU-MU strategy applies the threshold update technique to the MU method.
—TU-OU strategy applies the threshold update technique to the OU method.
—TU-SA strategy applies the threshold update technique to the SA method.
—TU-MEE strategy applies the threshold update technique to the MEE method.

5. EVALUATION

5.1 Experimental Settings

In this section, we analyze the effectiveness of four simple stopping criteria for active learning, including Maximum Uncertainty (MU), Overall Uncertainty (OU), Selected Accuracy (SA), and Minimum Expected Error (MEE), and four threshold update strategies: TU-MU, TU-OU, TU-SA, and TU-MEE. Because it is costly and practically infeasible to prepare a separate development dataset to determine an appropriate threshold for a stopping criterion, we use a common dataset in our experiments. We first test each stopping criterion with different threshold values such as {0.1, 0.01, 0.001, 0.0001} for MU, OU, and MEE, and {0.9, 0.95, 1.0} for SA, respectively.

[3]For example, an unlabeled example x was classified into class A at ith iteration, and class B at i+1th iteration.

Algorithm: Threshold Update Strategy

Given:

- A confidence-based stopping criterion *SC*: max-uncertainty or overall-uncertainty or selected-accuracy or minimum-expected-error
- The predefined threshold for the stopping criterion *SC* is initially set to β

Steps (during active learning process):

1) First check whether *SC* is satisfied. If yes, go to 2);
2) Then check whether classification-change criterion is satisfied. If yes, go to 4), otherwise go to 3);
3) Automatically update[4] the current threshold to be a new smaller value for max-uncertainty or overall-uncertainty or minimum-expected-error criterion, or to be a new larger value for selected-accuracy criterion, and then go to 1).
4) Stop active learning process.

Fig. 3. Threshold update algorithm.

In the experimental setting of each threshold update strategy, the initial threshold is equally set to 0.1 for MU, OU, and MEE, and 0.9 for SA. For the threshold revision during active learing process, the threshold value decreases by 0.01 each time for MU, OU, and MEE, and increases by 0.1 each time for SA, as shown in Figure 3.

In active learning, when the classifier firstly reaches the highest performance, it is suggested that the labeled data can sufficiently train a classifier with maximum effectiveness, and the active learning process can stop. We refer to such a time point as the *Best Stopping Time* (BST) point for ending the active learning process. That is, the best stopping criterion can make the active learning process stop at the BST point. To evaluate the effectiveness of each stopping criterion, we analyze the difference between the BST point and the stopping time point predicted by each stopping criterion. The smaller the difference between both time points, the better the stopping criterion. Therefore, a time point can be represented in the form of the number of unlabeled examples learned for human labeling. For notational convenience, the difference Δ_{SC} between the BST point and the stopping time point predicted by a stopping criterion *SC* is defined by

$$\Delta_{SC} = |\xi_{BST} - \xi_{PT}|,$$

where ξ_{BST} and ξ_{PT} denote the percentage of unlabeled examples in the pool U that have been learned at the BST and the Predicted Time (PT) points of the stopping criterion *SC*, respectively.

In the following active learning comparison experiments, each algorithm starts with a randomly chosen initial training set of 10 labeled examples, and makes 20 queries for each active learning iteration. A 10 by 10-fold cross-validation was performed. All results reported are the average of 10 trials in each active learning process. We utilize a Maximum Entropy (MaxEnt) model

[4]The threshold revision implementation is discussed in Section 5.1.

to design the basic classifier used in active learning. The advantage of the MaxEnt model is the ability to freely incorporate features from diverse sources into a single, well-grounded statistical model [Berger et al. 1996]. A publicly available MaxEnt toolkit[5] was used in this experiment. The *accuracy* was used as the performance metric in the following experiments.

5.2 Evaluation Datasets

Following previous studies on active learning for NLP applications [Lewis and Gale 1994; McCallum and Nigam 1998a; Tong and Koller 2001; Chen et al. 2006; Zhu and Hovy 2007; Chan and Ng 2007], to evaluate the effectiveness of each stopping criterion for pool-based active learning, we constructed some active learning experiments for three types of NLP applications including word sense disambiguation, text classification and opinion analysis tasks, using seven publicly available real-world datasets.

—*Word Sense Disambiguation Task*. Two publicly available real-world datasets are used in this task: *OntoNotes* and *Interest* datasets. The *OntoNotes* project [Hovy et al. 2006] uses the WSJ part of the Penn Treebank [Marcus et al. 1993]. The senses of noun words occurring in OntoNotes are linked to the Omega ontology [Philpot et al. 2005]. In this experiment, we focus on the 10 most frequent nouns used in our previous work [Zhu and Hovy 2007]: *rate*, *president*, *people*, *part*, *point*, *director*, *revenue*, *bill*, *future*, and *order*. The *Interest* dataset was developed by Bruce and Wiebe [1994]. It consists of 2369 sentences of the noun "*interest*" with its correct sense manually labeled. The noun "*interest*" has six different senses in this dataset. The interest dataset has been previously used for a WSD study [Ng and Lee 1996]. To build the MaxEnt-based classifier for the WSD task, three knowledge sources are used to capture contextual information: *unordered single words in topical context*, *POS of neighboring words with position information*, *and local collocations*. These are the same the knowledge sources used in other word sense disambiguation studies [Lee and Ng 2002].

—*Text Classification Task*. Four publicly available datasets are used in this active learning comparison experiment: *WebKB*, *Comp2a*, *Comp2b,* and *Comp2c* datasets. The WebKB dataset has been widely used in text classification research. Following previous studies [McCallum and Nigam 1998b], we use the four most populous categories: *student, faculty, course,* and *project*. The Comp2a dataset consists of *comp.os.ms-windows.misc* and *comp.sys.ibm.pc.hardware* subset of 20 news groups. The Comp2b dataset consists of *comp.graphics* and *comp.windows.x* categories from 20 news groups. The Comp2c dataset consists of *alt.atheism* and *talk.religion.misc* classes. These datasets have been previously used in active learning for text classification [Roy and McCallum 2001; Schein and Ungar 2007]. We first processed all datasets by running the corpus with a stop-word list. The MaxEnt model has been used to design the text classifier. No feature selection technique is used for the text classifier, because experimental results show that

[5]See http://homepages.inf.ed.ac.uk/s0450736/maxent_ toolkit.html.

Table I. Descriptions of These Datasets Used in Following Active Learning Evaluation (including: the number of classes and class distribution)

Dataset		Classes	Class Distribution
OntoNotes	rate	2	208/934/
	president	3	936/157/17/
	people	4	815/67/7/5/
	part	4	102/454/16/75/
	point	7	471/37/88/19/12/3/7/
	director	2	637/35/
	revenue	2	517/23/
	bill	4	349/122/40/3/
	future	3	23/82/409/
	order	8	342/6/61/54/4/2/6/3/
Interest		6	500/1252/178/66/361/11/
WebKB		4	504/930/1641/1124/
Comp2a		2	983/1000/
Comp2b		2	999/1000/
Comp2c		2	1000/1000/
MPQA		2	4958/6081/

using feature selection seems to have negative effects on the performance of active learning for text classification.

—*Opinion Analysis Task*. To analyze an opinioned text, the first step is to build a classifier that identifies opinion-bearing sentences in the text under a two-way classification framework [Kim 2006]. These opinion-bearing sentences express an opinion, namely subjective sentences. The Multi-Perspective Question Answering (MPQA) opinion corpus [Wiebe et al. 2003] contains news articles manually annotated using an annotation scheme for subjectivity. According to the opinion annotation scheme, all sentences in the MPQA can be divided into two categories: *subjective* and *objective*. As shown in Table I, the MPQA corpus contains 4958 objective sentences and 6081 subjective sentences. In this work, the opinion analysis is viewed as a binary classification task in which the MaxEnt model is also utilized to design the classifier for active learning, and only four types of tokens are considered as features, including *noun, verb, adjective,* and *adverb*.

5.3 Effectiveness Evaluation in Terms of Δ_{SC}

The first set of experiments conducted demonstrate the relative performance of several different stopping criteria on multiple datasets in terms of Δ_{SC}. Figure 4 shows the effectiveness of each stopping criterion with different thresholds on evaluation datasets in terms of the Δ_{SC} measure. The smaller the difference Δ_{SC} value is, the better the stopping criterion SC. $\Delta_{SC} = 0$ indicates that the BST point and the predicted time point are the same.

For MU, OU, and MEE, a smaller predefined threshold generally results in learning more unlabeled examples for human labeling. For SA, a larger threshold results in learning more unlabeled examples. As mentioned in Sections 4.1, 4.2, and 4.4, MU, OU, and MEE methods are preferable with a small threshold,

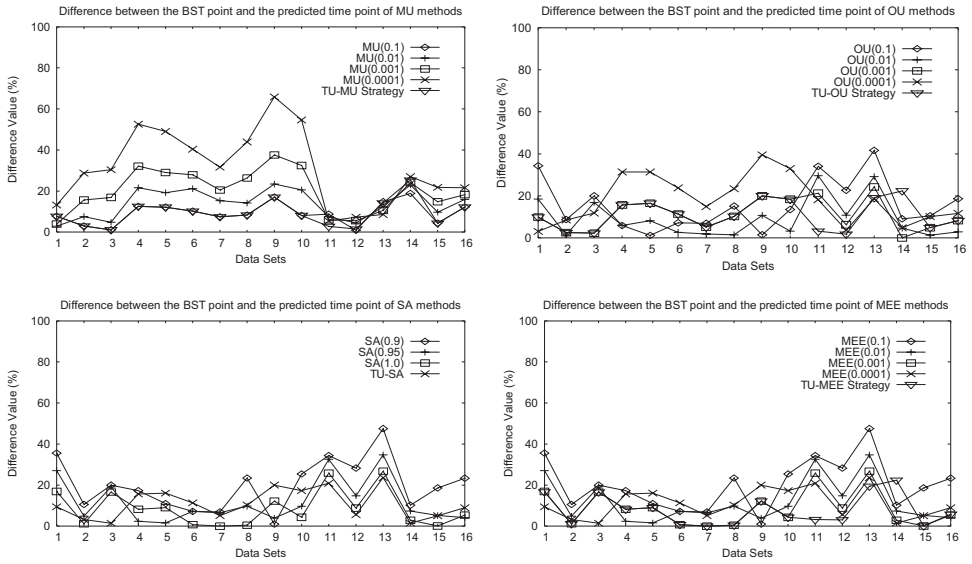

Fig. 4. Effectiveness of various stopping criteria with different thresholds. The numbers varying from 1–16 in the x-axis represent various evaluation datasets {*Rate, President, People, Part, Point, Director, Revenue, Bill, Future, Order, WebKB, Comp2a, Comp2b, Comp2c, Interest, MPQA*}, respectively. The difference value (%) in the y-axis denotes the difference Δ_{SC} between the BST point and the time point predicted by a stopping criterion *SC*.

because a classifier with maximum effectiveness can result in sufficient confidence and the lowest expected error on its classification of the remaining unlabeled data. Take MEE for example, MEE(0.0001) outperforms MEE(0.1) on most evaluation datasets except on the 5th, 6th, and 9th datasets (*Point, Director, Future*). For MU methods, on the other hand, among various MU methods with different thresholds, MU(0.0001) is the worst, and MU(0.1) works the best. In this case, a smaller threshold seems to result in worse performance for MU.

Seen from the preceding right-top and left-bottom figures in Figure 4, a fixed threshold cannot guarantee that an individual confidence-based stopping criterion can obtain satisfactory performance on all datasets. It is a crucial issue how to choose an appropriate threshold parameter for an individual stopping criterion in a specific application. The threshold update strategy was thus designed to overcome this problem. Actually the advantage of the threshold update strategy is to consider an additional classification label change of each unlabeled example during the active learning process. Figure 4 shows that by dynamically adjusting the predefined threshold for a stopping criterion on each evaluation dataset, the threshold update technique can improve the performance of each individual stopping criterion on most datasets. The threshold update technique makes each individual stopping criterion feasible in a specific application, because the strict requirement of a fixed predefined threshold is not needed.

Table II. Average Difference Δ_{ave} Values of Each Stopping Criterion Over All Evaluation Datasets, Compared with the BST Point

Methods	MU(0.1)	MU(0.01)	MU(0.001)	MU(0.0001)	TU-MU
$\Delta_{ave}(\%)$	9.2	13.9	20.1	31.4	9.2
Methods	OU(0.1)	OU(0.01)	OU(0.001)	OU(0.0001)	TU-OU
$\Delta_{ave}(\%)$	15.7	9.3	11.1	17.9	10.7
Methods	SA(0.9)	SA(0.95)	SA(1.0)	TU-SA	
$\Delta_{ave}(\%)$	10.3	14.1	15.5	10.2	
Methods	MEE(0.1)	MEE(0.01)	MEE(0.001)	MEE(0.0001)	TU-MEE
$\Delta_{ave}(\%)$	20.1	11.9	8.7	10.9	**7.7**

The smaller the difference Δ_{ave} value is, the better the stopping criterion *SC* is.

To further provide a quantitative analysis of each stopping criterion, Table II shows average difference Δ_{ave} values of each stopping criterion over all evaluation datasets. A smaller Δ_{ave} value indicates the stopping criterion *SC* achieves a better performance, in comparison to the BST point.

As shown in Table II, for SA, MEE, and MU, the threshold update strategy also achieves the best performance among the corresponding methods with different thresholds. The threshold update strategy TU-MEE achieves the best performance of 7.7% Δ_{ave} value among all methods. TU-OU outperforms OU methods except OU(0.01), because of a possible reason that TU-OU obtains unsatisfactory performance on the 14th dataset (*Comp2c*). From experimental results, we find that on the Comp2c dataset, the classifier has achieved the highest performance by learning 14.21% of examples, but the classification labels of the remaining unlabeled examples are still greatly changing at the corresponding time point of the active learning process. In such case, the threshold update strategy would allow active learning to continue.

To further explore the effectiveness of each confidence-based stopping criterion, we analyze the performance of each stopping criterion on the evaluation dataset under various ΔSC constraint conditions. The number of evaluation datasets on which each stopping criterion works well under various Δ_{SC} constraint conditions is listed in each cell of Table III.

TU-MEE and MEE(0.001) achieve Δ_{SC} values less than 0.5% on only three datasets, and less than 1% on four datasets. Among various MU methods, only MU(0.1) achieves an Δ_{SC} value less than 1% on one dataset, and none for the constraint of the Δ_{SC} value less than 0.5%. It is noteworthy that TU-MEE and OU(0.01) obtain Δ_{SC} values less than 5% on 8 out of 16 sets, and all threshold update strategies can obtain Δ_{SC} values less than 10% on half of the datasets. Under the constraint condition of $\Delta_{SC} < 10\%$, TU-SA and SA(0.9) can work well on 12 out of 16 sets (i.e., 3/4 datasets). Although OU(0.01) can slightly outperform TU-OU as shown in Table III, we cannot say that 0.01 is the appropriate threshold value for OU in any given task. Table III shows that in most cases the threshold update technique can improve each confidence-based stopping criterion. As mentioned before, the advantage of the threshold update technique can solve the problem of the fixed predefined threshold in real-world applications.

Table III. Analysis of the Effectiveness of Each Stopping Criterion in Terms of Δ_{SC} Measure

Δ_{SC}	<0.5%	<1%	<3%	<5%	<10%	<15%	<20%	≤100%
MU(0.1)		1	3	4	9	**14**	**16**	16
MU(0.01)			1	3	6	8	11	16
MU(0.001)				1	3	5	8	16
MU(0.0001)					3	4	4	16
TU-MU			4	5	9	**14**	15	16
OU(0.1)			2	2	7	9	12	16
OU(0.01)			**6**	**8**	10	12	14	16
OU(0.001)	1	1	3	4	8	10	14	16
OU(0.0001)				3	4	8	10	16
TU-OU			3	5	8	10	15	16
SA(0.9)		1	4	6	**12**	**14**	14	16
SA(0.95)	1	1	1	3	8	13	13	16
SA(1.0)	1	1	1	2	5	13	13	16
TU-SA	1	2	5	6	**12**	**14**	14	16
MEE(0.1)		1	1	1	3	6	9	16
MEE(0.01)			2	5	11	12	13	16
MEE(0.001)	**3**	**4**	**6**	7	11	12	14	16
MEE(0.0001)			2	3	8	10	14	16
TU-MEE	**3**	**4**	5	**8**	11	12	15	16

Each digital number in a row indicates the number of evaluation datasets on which the difference Δ_{SC} between the predicted time point and the BST point is satisfied with the constraint condition shown in the corresponding column. The bold number indicates the best performance under current Δ_{SC} constraints.

5.4 Effectiveness Evaluation in Terms of ACC_Δ_{SC}

To further analyze the effectiveness of each stopping criterion, here we focus on the differences between accuracies obtained at the BST point and the predicted stopping time point of each stopping criterion, namely the accuracy difference ACC_Δ_{SC}. The accuracy difference ACC_Δ_{SC} between the BST point and the predicted stopping time point of a stopping criterion SC is defined by

$$ACC\_\Delta_{SC} = ACC_{BST} - ACC_{PT},$$

where ACC_{BST} and ACC_{PT} denote the classifier's accuracy performance obtained at the BST and the Predicted stopping Time (PT) points of a stopping criterion SC, respectively. The value $ACC\_\Delta_{SC} > 0$ indicates that the accuracy obtained at the predicted stopping time point is higher than that of the BST point. Similarly, $ACC\_\Delta_{SC} < 0$ indicates the worse case, and $ACC\_\Delta_{SC} = 0$ indicates that the BST point and the predicted stopping time point are the same.

The second set of experiments conducted evaluate the relative performance of several different stopping criteria on multiple datasets in terms of the accuracy difference ACC_Δ_{SC}, as shown in Figure 5 and Table IV. We think that a good stopping criterion can make an active learning process stop at the BST point as close as possible, and can achieve the same or higher performance (i.e., $ACC\_\Delta_{SC} = 0$ or $ACC\_\Delta_{SC} > 0$), comparing to that obtained at the BST point.

Figure 5 depicts that the MU and SA methods achieve accuracies close to that of the corresponding BST points on most datasets. On the 14th set (*Comp2c*), MU methods achieve slightly better performance than that of the corresponding BST points. OU(0.1) and MEE(0.1) obviously obtain unsatisfactory performance on the 11th–16th datasets (*WebKB, Comp2a, Comp2b, Comp2c, Interest*

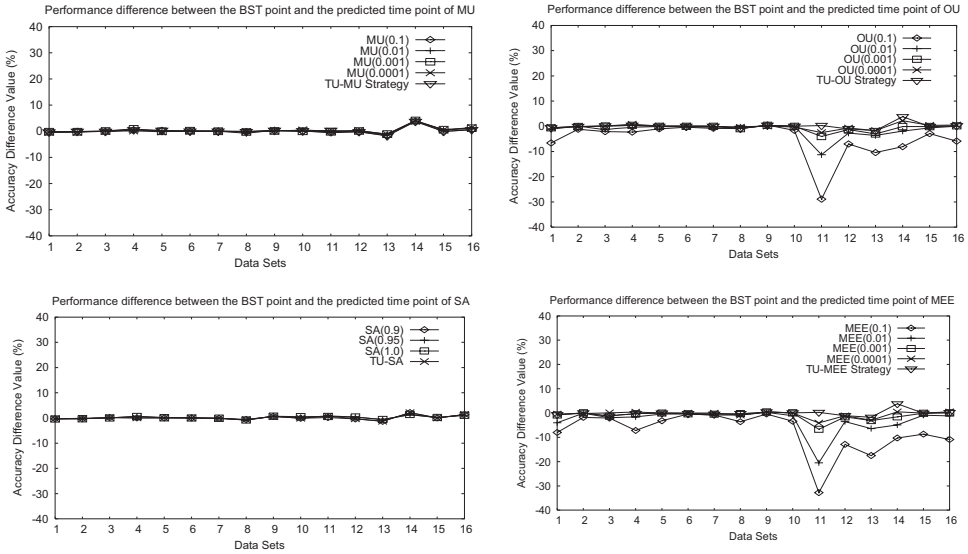

Fig. 5. Accuracy effectiveness of various stopping criteria with different thresholds. The numbers varying from 1–16 in the x-axis represent various evaluation datasets {*Rate, President, People, Part, Point, Director, Revenue, Bill, Future, Order, WebKB, Comp2a, Comp2b, Comp2c, Interest, MPQA*}, respectively. Y-axis denotes the difference values ACC_Δ_{SC} between the accuracies obtained at BST point and the predicted stopping time point of a stopping criterion *SC*.

Table IV. Average Accuracy Difference ACC_Δ_{ave} Values of Each Stopping Criterion over All Evaluation Datasets, Compared with the BST Point

Methods	MU(0.1)	MU(0.01)	MU(0.001)	MU(0.0001)	TU-MU
ACC_Δ_{ave}(%)	0.15	0.21	0.33	0.29	0.19
Methods	OU(0.1)	OU(0.01)	OU(0.001)	OU(0.0001)	TU-OU
ACC_Δ_{ave}(%)	−4.97	−1.48	−0.56	−0.13	0.05
Methods	SA(0.9)	SA(0.95)	SA(1.0)	TU-SA	
ACC_Δ_{ave}(%)	0.08	0.20	0.21	0.12	
Methods	MEE(0.1)	MEE(0.01)	MEE(0.001)	MEE(0.0001)	TU-MEE
ACC_Δ_{ave}(%)	−7.75	−3.01	−0.98	−0.49	−0.07

and *MPQA*). It is noteworthy to observe that these threshold update strategies achieve average accuracy performances very close to the accuracy performances obtained at the corresponding BST points, and sometimes a better performance ($ACC\_\Delta_{ave} > 0$) is achieved by the threshold update method.

Table IV shows the average accuracy difference ACC_Δ_{ave} values of each stopping criterion over all evaluation datasets, compared with the BST point. Except OU(0.1), MEE(0.1), OU(0.01), and MEE(0.01), other methods can achieve accuracies within less than 1% difference to the highest accuracies obtained at the BST points. Compared with the BST points, TU-MU, TU-OU, and TU-SA achieve slightly better performances while TU-MEE obtains a −0.07% ACC_Δ_{ave} value. As mentioned earlier, a good stopping criterion should achieve the same or higher performance comparing to that of the BST point. From this

perspective, except for OU(0.1), MEE(0.1), OU(0.01), and MEE(0.01), there is no significant difference among these methods in terms of average accuracy difference ACC_Δ_{ave}. Table IV shows that all threshold update techniques achieve satisfactory performance in terms of ACC_Δ_{ave}, in comparison to the performance of BST points.

6. RELATED WORK

The most similar work done in Vlachos [2008] is to define a stopping criterion of active learning based on the estimate of classifier's confidence, in which a separate outside dataset is used to estimate the classifier's confidence. The key issue is how to identify a consistent drop in the confidence of a classifier during an active learning process. This method is simialr to our OU method. Our OU method estimates the classifier's confidence on the remaining unlabeled data rather than on an outside dataset. Like our OU method, Vlachos [2008] considered average entropy as the confidence measure for active learning with a probabilistic classifier such as MaxEnt, and the average margin for active learning with a nonprobabilistic classifier such as SVMs. Sometimes it is difficult to sufficiently identify the ideal drop in the confidence of a classifier during the active learning process in real-world applications [Laws and Schutze 2008]. For such case, a local maximum of the confidence curve would be found.

Maximizing expected error reduction [Roy and McCallum 2001] is often used as a example selection criterion in the pool-based active learning setting. Campbell et al. [2000], Roth and Small [2008], Donmez et al. [2007], and Dimitrakakis and Savu-Krohn [2008] independently applied a stopping criterion based on the estimation of the probability of error on the remaining unlabeled data in their tasks, similar to our MEE method.

Schohn and Cohn [2000] proposed a stopping criterion for active learning with support vector machines based on an assumption that the data used is linearly separable. However, in most real-world cases this assumtion seems to be a little unreasonable and difficult to satisfy. Also, their stopping criterion cannot be applied for active learning with other types of classifiers such as NB and MaxEnt models.

Tomanek et al. [2007] considered a factor of the disagreement rate between the classifiers to predict the stopping point for active learning with committee-based sampling. This method assumes the classifier has strong confidence on its classification on the reamining unlabeled examples if the disagreement rate is close to zero. This method is similar to the classification label change factor used in our threshold update strategy.

Laws and Schutze [2008] presented a gradient-based stopping criterion for active learning of named entity recognition. In their method, the active learning stops if the gradient of the performance curve approaches 0. The key issue is how to sufficiently estimate the gradient of the performance curve during the active learning process in real-world applications. As shown in Figure 2, increasing of the performnce is often not monotonic. For such a case, setting an appropriate window size for gradient estimation is crucial in a specific

task. Otherwise, active learning would stop at the local maximum point of the performance curve.

7. DISCUSSION

In this section, we will discuss the possibility of applying our stopping criteria for other variations of active learning settings. Among our four stopping criteria, SA can be directly used for active learning with probabilistic or non-probabilistic classifiers. But SA can only be applied to batch mode active learning, because SA is based on the feedback from Oracle, and too few uncertain candidates in each learning cycle may not be adequate for obtaining a reasonable feedback. For MU and OU, the key is how to measure the uncertainty of each unlabeled example in other variations of active learning settings. The crucial issue of our MEE is how to estimate the expected error on the unlabeled examples in other variations of active learning settings.

An alternative method of uncertainty sampling is to pick the most uncertain unlabeled example with the smallest margin. The margin is calculated as the difference between the largest two class probabilities produced by the probabilistic classifier. In this case, the uncertainty measurement function $UM(.)$ used by MU and OU can be defined by means of the margin instead of the entropy. Tong and Koller [2001] presented a way of performing uncertainty sampling with SVMs by using the decision margin of the classifier. The closer a datapoint lies to the hyperplane, the more uncertain it seems to be. Since SVMs do not yield probabilistic output but a decision margin, the sigmoid function [Platt 1999] can be used to obtain the probabilistic outputs which are needed in defining MU, OU, and MEE stopping criteria for the setting of active learning with SVMs.

In the ensemble-based active learning setting, the regions of uncertain classification are often where the classifiers give different answers. To define an MEE stopping criterion for such a case, we can adopt a method to estimate the final class posterior for an unlabeled example as the unweighted average of the class posteriors for each of the classifiers, as used in Roy and McCallum [2001]. This bagged posterior is more reflective of the true uncertaity [Roy and McCallum 2001]. Korner and Wrobel [2006] applied four different techniques to measure ensemble disagreement such as *margin-based disagreement*, *uncertainty sampling-based disagreement*, *entropy-based disagreement*, and *specific disagreement ("control")*. To apply our MU and OU stopping criteria for ensemble-based active learning, the uncertainty of each unlabeled example can be estimated by means of these ensemble disagreement measures [Korner and Wrobel 2006].

Vlachos [2008] presented a stopping criterion of active learning in which a separate and large outside dataset is required in advance to estimate the classifier's confidence, and feature extraction needs to be performed. The confidence is estimated in terms of the average margin for active learning with SVMs, and the average entropy for active learning with a maximum-entropy-based classifier. In our work, the confidence estimation for each confidence-based stopping criterion is done within the unlabeled pool U during the active learning process.

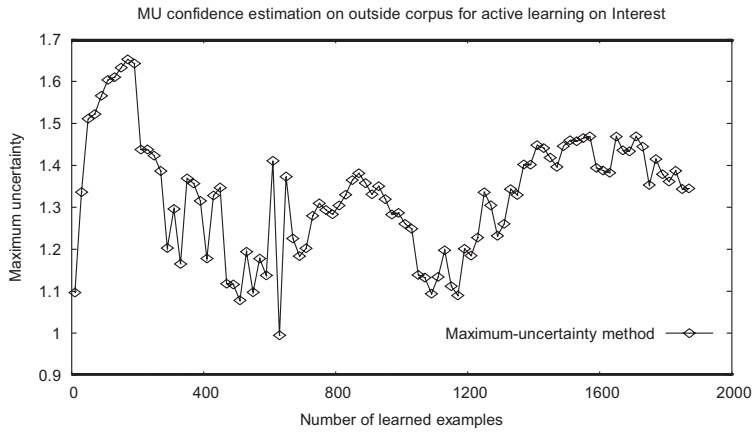

Fig. 6. MU's confidence estimation on outside corpus for active learning with uncertainty sampling on Interest dataset. In the MU method, the maximum uncertainty stands for the entropy of the most uncertain examples chosen at each learning cycle.

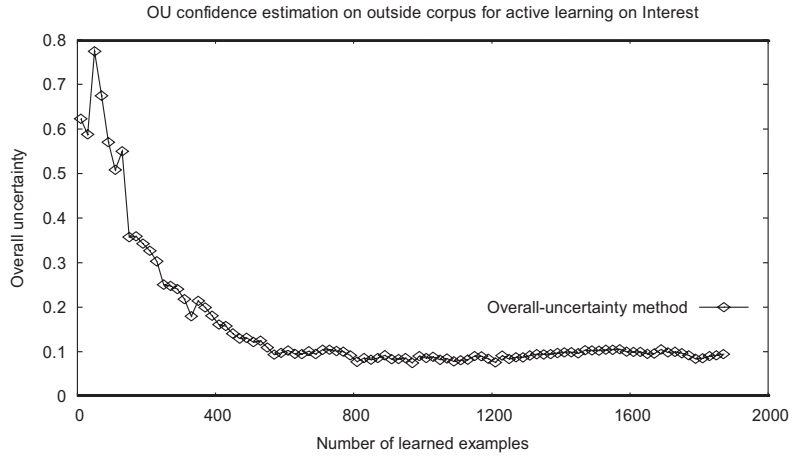

Fig. 7. OU's confidence estimation on outside corpus for active learning with uncertainty sampling on Interest dataset. In the OU method, overall uncertainty stands for the average entropy of all remaining unlabeled examples in the pool U at each learning cycle.

We also tried to evaluate the confidence of a classifier using an outside unlabeled corpus for each confidence-based stopping criterion except the SA method. This is because the SA method does not rely on the outside unlabeled dataset. To investigate the impact of confidence estimation (on outside unlabeled corpus) on the effectiveness of each confidence-based stopping criterion, here we give some exemplary figures as in Figures 6 through 8.

Figures 6, 7, and 8 depict the confidence estimation on outside corpus for each confidence-based stopping criterion such as MU, OU, or MEE during an active learning process. We can see from Figure 6 that MU cannot work in this

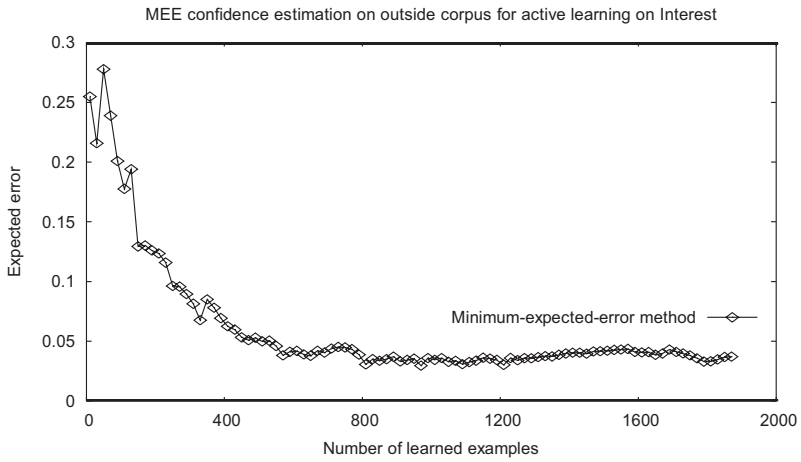

Fig. 8. MEE's confidence estimation on outside corpus for active learning with uncertainty sampling on Interest dataset. The expected errors are calculated on all remaining unlabeled examples in the pool U at each learning cycle.

case. Figures 7 and 8 show that the curves for OU and MEE lay out a decreasing trend in that both curves become almost flat after iteration 600. This shows the possibility of applying our OU and MEE methods, for which a separate outside unlabeled corpus is used for confidence estimation. From this perspective, the stopping criterion defined in Vlachos [2008] is similar to our OU method.

8. CONCLUSION AND FUTURE WORK

In this article, we address the stopping criterion issue of active learning, and analyze the problems faced by some common approaches to stopping the active learning process. We further present four simple confidence-based stopping criteria, including maximum uncertainty, overall uncertainty, selected accuracy, and minimum expected error, to determine when to stop an active learning process. To solve the problem of obtaining a proper threshold for each confidence-based stopping criterion in a specific task, a threshold update strategy is presented by considering the classification label change of unlabeled examples during the active learning process. In this strategy the predefined threshold of each stopping criterion can be automatically adjusted during the active learning process. The effectiveness of these proposed stopping criteria for active learning is evaluated on three NLP tasks, namely word sense disambiguation, text classification, and opinion analysis, using seven real-world datasets. Some interesting future work is to investigate further how to combine the best of these criteria, and how to consider performance change to define an appropriate stopping criterion for active learning.

ACKNOWLEDGMENTS

We would like to thank Muhua Zhu and Tong Xiao for helpful discussion, and the anonymous reviewers for valuable comments and corrections.

REFERENCES

ANGLIUN, D. 1988. Queries and concept learning. *Mach. Learn. 2*, 3, 319–342.

BARAM, Y., EL-YANIV, R., AND LUZ, K. 2004. Online choice of active learning algorithms. *J. Mach. Learn. Res. 5*, 255–291.

BECKER, M. AND OSBORNE, M. 2005. A two-stage method for active learning of statistical grammars. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence.* 991–996.

BERGER, A. L., DELLA PIETRA, S. A., AND DELLA PIETRA, V. J. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist. 22*, 1, 39–71.

BRUCE, R. AND WIEBE, J. 1994. Word-Sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. 139–146.

CAMPBELL, C., CRISTIANINI, N., AND SMOLA, A. 2000. Query learning with large margin classifiers. In *Proceedings of the International Conference on Machine Learning*. 111–118.

CHAN, Y. S. AND NG, H. T. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 49–56.

CHEN, J., SCHEIN, A., UNGAR, L., AND PALMER, M. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. 120–127.

COHN, D. A., ATLAS, L., AND LADNER, R. E. 1994. Improving generalization with active learning. *Mach. Learn. 15*, 201–221.

COHN, D. A., GHAHRAMANI, Z., AND JORDAN, M. I. 1996. Active learning with statistical models. *J. Artif. Intell. Res. 4*, 129–145.

CULOTTA, A. AND MCCALLUM, A. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05).* 746–751.

DIMISTAKAKIS, C. AND SAVU-KROHN, C. 2008. Cost-Minimizing strategies for data labeling: Optimal stopping and active learning. In *Proceedings of the 5th International Symposium on Foundations of Information and Knowledge Systems (FoIKS)*. 96–111.

DONMEZ P., CARBONELL, J. G., AND BENNETT, P. N. 2007. Dual strategy active learning. In *Proceedings of the European Conference on Machine Learning (ECML)*. 1–12.

DUDA, R. O. AND HART, P. E. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York.

DAGAN, I. AND ENGELSON, S. P. 1995. Committee-Based sampling for training probabilistic classifiers. In *Proceedings of the International Conference on Machine Learning*. 150–157.

FREUND, Y., SEUNG, H. S., SHAMIR, E., AND TISHBY, N. 1997. Selective sampling using the query by committee algorithm. *Mach. Learn. 28*, 2, 133–168.

HOVY, E. H., MARCUS, M., PALMER, M., RAMSHAW, L., AND WEISCHEDEL, R. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*. 57–60.

HWA, R. 2000. Sample selection for statistical grammar induction. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. 45–52.

KORNER, C. AND WROBEL, S. 2006. Multi-Class ensemble-based active learning. In *Proceedings of the European Conference on Machine Learning (ECML).* 687–694.

KIM, S. M. 2006. Identification, classification, and analysis of opinions on the Web. Ph.D. thesis, University of Southern California.

JONES, R. 2005. Learning to extract entities from labeled and unlabeled text. Ph.D. thesis, Carnegie Mellon University.

LAWS, F. AND SCHÜTZE, H. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*. 465–472.

LEE, Y. AND NG, H. 2002. An empirical evaluation of knowledge sources and learning algorithm for word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 41–48.

LEWIS, D. AND GALE, W. 1994. A sequential algorithm for training text classifiers. In *Proceedings of 17th ACM International Conference on Research and Development in Information Retrieval*. 3–12.

LI, M. AND SETHI, I. K. 2006. Confidence-Based *active learning. IEEE Trans. Patt. Anal. Mach. Intell. 28*, 8, 1251–1261.

MARCUS, M., SANTORINI, B., AND MARCINKIEWICZ, M. 1993. Building a large annotated corpus of English: the Penn treebank. *Comput. Linguist. 19*, 2, 313–330.

MCCALLUM, A. AND NIGRAM, K. 1998a. Employing EM in pool-based active learning for text classification. In *Proceedings of 15th International Conference on Machine Learning*. 350–358.

MCCALLUM, A. AND NIGRAM, K. 1998b. A comparison of event models for naïve Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*.

NG, H. AND LEE, H. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics*. 40–47.

NGAI, G. AND YAROWSKY, D. 2000. Rule writing or annotation: Cost-Efficient resource usage for based noun phrase chunking. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics*. 117–125.

PHILPOT, A., HOVY, E. H., AND PANTEL, P. 2005. The Omega ontology. In *Proceedings of OntoLex Conference onOntologies and Lexical Resources*. 59–66.

PLATT, J. 1999. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Classifiers*. 61–74.

ROTH, D. AND SMALL, K. 2008. Active learning for pipeline models. In *Proceedings of the National Conference on Artificial Intelligence*. 683–688.

ROY, N. AND MCCALLUM, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning*. 441–448.

SCHEIN, A. AND UNGAR, L. 2007. Active learning for logistic regression: An evaluation. *Mach. Learn. 68*, 235–265.

SCHOHN, G. AND COHN, D. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the 17th International Conference on Machine Learning*. 839–846.

SEUNG, H. S., OPPER, M., AND SOMPOLINSKY, H. 1992. Query by committee. In *Proceedings of the 5th Annual ACM Conference on Computational Learning Theory*. 287–294.

SHEN, D., ZHANG, J., SU, J., ZHOU, G., AND TAN, C. 2004. Multi-Criteria-Based active learning for named entity recognition. In *Proceedings of the 42th Annual Meeting of the Association of Computational Linguistics*.

TANG, M., LUO, X., AND ROUKOS, S. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*. 120–127.

TONG, S. AND KOLLER, D. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res. 2*, 45–66.

TOMANEK, K., WERMTER, J., AND HAHN, U. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning*. 486–495.

THOMPSON, C. A., CALIFF, M. E., AND MOONEY, R. J. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning*. 406–414.

VLACHOS, A. 2008. A stopping criterion for active learning. *Comput. Speech Lang. 22*, 3, 295–312.

WIEBE, J., BRECK, E., BUCKLEY, C., CARDIE, C., DAVIS, P., ET AL. 2003. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*.

ZHU, J. AND HOVY, E. H. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the Joint Meeting of the Conference on*

Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning. 783–790.

ZHU, J., WANG, H., AND HOVY, E. H. 2008a. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*. 366–372.

ZHU, J., WANG, H., AND HOVY, E. H. 2008b. Multi-Criteria-Based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics*. 1129–1136.

Received July 2008; revised May 2009; accepted February 2010