Multi-Criteria-based Strategy to Stop Active Learning for Data Annotation

Jingbo Zhu Huizhen Wang
Natural Language Processing Laboratory
Northeastern University
Shenyang, Liaoning, P.R.China 110004
zhujingbo@mail.neu.edu.cn
wanghuizhen@mail.neu.edu.cn

Eduard Hovy
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
hovy@isi.edu

Abstract

In this paper, we address the issue of deciding when to stop active learning for building a labeled training corpus. Firstly, this paper presents a new stopping criterion, classification-change, which considers the potential ability of each unlabeled example on changing decision boundaries. Secondly, a multi-criteria-based combination strategy is proposed to solve the problem of predefining an appropriate threshold for each confidence-based stopping criterion, such as max-confidence, min-error, and overall-uncertainty. Finally, we examine the effectiveness of these stopping criteria on uncertainty sampling and heterogeneous uncertainty sampling for active learning. Experimental results show that these stopping criteria work well on evaluation data sets, and the combination strategies outperform individual criteria.

1 Introduction

Creating a large labeled training corpus is very expensive and time-consuming in some real-world applications. For example, it is a crucial issue for automated word sense disambiguation task, because validations of sense definitions and sense-tagged data annotation have to be done by human experts, e.g. OntoNotes project (Hovy *et al.,* 2006).

Active learning aims to minimize the amount of human labeling effort by automatically selecting the most informative unlabeled example for human annotation. In recent years active learning has been widely studied in natural language processing (NLP) applications, such as word sense disambiguation (WSD) (Chen *et al.,* 2006; Zhu and Hovy, 2007), text classification (TC) (Lewis and Gale, 1994; McCallum and Nigam, 1998a), named entity recognition (Shen *et al.,* 2004), chunking (Ngai and Yarowsky, 2000), and statistical parsing (Tang *et al.*, 2002).

However, deciding when to stop active learning is still an unsolved problem and seldom mentioned issue in previous studies. Actually it is a very important practical issue in real-world applications, because it obviously makes no sense to continue the active learning procedure until the whole unlabeled corpus has been labeled. The active learning process can be ended when the current classifier reaches the maximum effectiveness. In principle, how to learn a stopping criterion is a problem of estimation of classifier (i.e. learner) effectiveness during active learning (Lewis and Gale, 1994).

In this paper, we address the issue of a stopping criterion for pool-based active learning with uncertainty sampling (Lewis and Gale, 1994), and propose a multi-criteria-based approach to determining when to stop active learning process. Firstly, this paper makes a comprehensive analysis on some confidence-based stopping criteria (Zhu and Hovy, 2007), including max-confidence, min-error and overall-uncertainty, then proposes a new stopping criterion, *classification-change*, which considers the potential ability of each unlabeled example on changing decision boundaries. Secondly, a combination strategy is proposed to solve the problem of predefining an appropriate threshold for each confidence-based stopping criterion in a specific task.

In uncertainty sampling scheme, the most uncertain unlabeled example is considered as the most informative case selected by active learner at each learning cycle. However, an uncertain example for one classifier may be not an uncer-

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1129–1136
Manchester, August 2008

tain example for other classifiers. When using active learning for real-world applications such as WSD, it is possible that a classifier of one type selects samples for training a classifier of another type, called the heterogeneous approach (Lewis and Catlett, 1994). For example, the final trained classifier for WSD is often different from the classifier used in active learning for constructing the training corpus.

To date, no one has studied the stopping criterion issue for the heterogeneous approach. In this paper, we examine the effectiveness of each stopping criterion on both traditional uncertainty sampling and heterogeneous uncertainty sampling for active learning. Experimental results of active learning for WSD and TC tasks show that these proposed stopping criteria work well on evaluation data sets, and the combination strategies outperform individual criteria.

2 Active Learning Process

In this paper, we are interested in *uncertainty sampling* for pool-based active learning (Lewis and Gale, 1994), in which an unlabeled example x with maximum uncertainty is selected to augment the training data at each learning cycle. The maximum uncertainty implies that the current classifier has the least confidence on its classification of this unlabeled example.

Actually active learning is a two-stage process in which 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 and retraining is adopted.

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. Use uncertainty sampling technique to select m most informative unlabeled examples, and ask oracle H for labeling
3. Augment L with these m new examples, and remove them from U
4. Use L to retrain the current classifier C

Until the predefined stopping criterion SC is met.

Figure 1. Active learning with uncertainty sampling technique

3 Stopping Criteria for Active Learning

In this section, we mainly address the problem of general stopping criteria for active leanring, and study how to define a reasonable and appropriate stopping criterion *SC* shown in Fig. 1.

3.1 Effectiveness Estimation and Confidence Estimation

To examine whether the classifier has reached the maximum effectiveness during active learning procedure, it seems an appealing solution when repeated learning cycles show no significant performance improvement. However, this is often not feasible. To investigate the impact of performance change on defining a stopping criterion for active learning, we first give an example of active learning for WSD shown in Fig. 2.

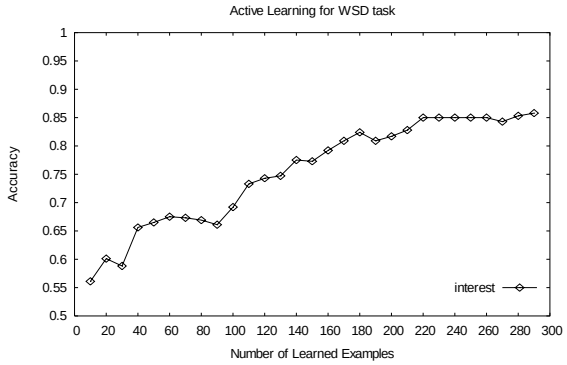

Figure 2. An example of active learning for WSD on word "*interest*".

Fig. 2 shows that the accuracy performance generally increases, but apparently degrades at iterations 30, 90 and 190, and does not change anymore during iterations 220-260 in the active learning process. Actually the first time of the highest performance of 91.5% is achieved at 900 which is not shown in Fig. 2. Although the accuracy performance curve shows an increasing trend, it is not monotonically increasing. It is not easy to automatically determine the point of no significant performance improvement on the validation set, because points such as 30 or 90 would mislead a final judgment.

Besides, there is a problem of performance estimation of the current classifier during active learning process, because a separate validation set should be prepared in advance, a procedure that causes additional (high) cost since it is often done manually. Besides, how many samples are required for the pregiven separate validation set is an open question. Too few samples may not be adequate for a reasonable estimate and may result in an incorrect result. Too many samples would increase the building cost.

To define a stopping criterion for active learning, Zhu and hovy (2007) considered the estimation of the classifier's effectiveness as the second

task of confidence estimation of the classifier on its classification of all remaining unlabeled data. In the following section, we first introduce two confidence-based criteria, max-confidence and min-error, proposed by Zhu and Hovy (2007).

3.2 Max-Confidence

In uncertainty sampling scheme, if the uncertainty value of the most informative unlabeled example is sufficiently small, we can assume that the current classifier has sufficient confidence on its classification of the remaining unlabeled data. So the active learning process can be ended. Based on such assumption, Zhu and Hovy (2007) proposed *max-confidence* criterion based on the uncertainty estimation of the most informative unlabeled example. Its strategy is to consider whether the uncertainty value of the most informative unlabeled example is less than a very small predefined threshold.

3.3 Min-Error

As shown in Fig. 1, in uncertainty sampling scheme, the current classifier has the least confidence on its classification of these top-m selected unlabeled examples. If the current classifier can correctly classify these most informative examples, we can assume that the current classifier have sufficient confidence on its classification of the remaining unlabeled data. Based on such assumption, Zhu and Hovy (2007) proposed *min-error* criterion based on feedback from the oracle. Its strategy is to consider whether the current classifier can correctly predict the labels on these selected unlabeled examples, or the accuracy performance of the current classifier on these most informative examples is larger than a predefined threshold.

3.4 Overall-Uncertainty

The motivation behind the *overall-uncertainty* method is similar to that of the max-confidence method. However, the max-confidence method only considers the most informative 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. Based on such assumption, we propose overall-uncertainty method which is to consider whether the average uncertainty value of all remaining unlabeled examples is less than a very small predefined threshold.

3.5 Classification-Change

There is another problem of estimating classifier performance during active learning process. Cross-validation on the training set is almost impractical during the active learning procedure, because the alternative of requiring a held-out validation set for active learning is counterproductive. Hence we should look for a self-contained method.

Actually the motivation behind uncertainty sampling is to find some unlabeled examples near decision boundaries, and use them to clarify the position of decision boundaries. The current classifier considers such unlabeled examples near decision boundaries as the most informative examples in uncertainty sampling scheme for active learning. In other words, we assume that an unlabeled example with maximum uncertainty has the highest chance to change the decision boundaries.

Based on the above analysis, we think the active learning process can stop if there is no unlabeled example that can potentially change the decision boundaries. However, in practice, it is almost impossible to exactly recognize 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 solve this problem, we make an assumption that labeling an unlabeled example may shift the decision boundaries if this example was previously "outside" and is now "inside". In other words, if an unlabeled example is automatically assigned to two different labels during two recent learning cycles[2], we think that the labeling of this unlabeled example has a good chance to change the decision boundaries.

Based on such assumption, we propose a new approach based on classification change of each unlabeled example during two recent consecutive learning cycles ("previous" and "current"), called the *classification-change* method. Its strategy is to stop the active learning process by considering whether no classification change happens to the remaining unlabeled examples during two recent consecutive learning cycles. If true, we assume that the current classifier has sufficient confidence on its classification of the remaining unla-

[2] For example, an unlabeled example x was classified into class A at i^{th} iteration, and class B at $i+1^{th}$ iteration.

beled data, because all unlabeled examples near decision boundaries have been exhausted, and no further labeling will affect active learner.

4 Combination Strategy

As for the above three confidence-based stopping criteria such as max-confidence, min-error and overall-uncertainty, how to automatically determine an appropriate threshold in a specific task is a crucial problem. We think that different appropriate thresholds are needed for various active learning applications.

To solve this problem, in this section we propose a general combination strategy by considering the best of both classification-change and a confidence-based criterion, in which the predefined threshold of the confidence-based stopping criterion can be automatically updated during active learning.

The motivation behind the general combination strategy is to check whether the active learning becomes stable (i.e. check whether the classification-change method is met) when the current confidence-based stopping criterion is satisfied. If not, we think there are some remaining unlabeled examples that can potentially shift the decision boundaries, even if they are considered as certain cases from the current classifier's viewpoints. In this case, the threshold of the current confidence-based stopping criterion should be automatically revised to keep continuing the active learning process. The general combination strategy can be summarized as follows.

Procedure: General combination strategy
Given:
- stopping criterion 1: max-confidence or min-error or overall-uncertainty
- Stopping criterion 2: classification-change
- The predefined threshold for stopping criterion 1 is initially set to β

Steps(during active learning process):
1. First check whether stopping criterion 1 is satisfied. If yes, go to 2;
2. Then check whether stopping criterion 2 is satisfied. If yes, goto 4), otherwise goto 3;
3. Automatically update the current threshold to be a new smaller value for max-confidence and overall-uncertainty, or to be a new larger value for min-error, and then goto 1.
4. Stop active learning process.

Figure 3. General combination strategy

- **Strategy 1**: This strategy combines the max-confidence and classification-change methods simultaneously.
- **Strategy 2**: This strategy combines the min-error and classification-change methods simultaneously.
- **Strategy 3**: This strategy combines the overall-uncertainty and classification-change methods simultaneously.

5 Evaluation

5.1 Experimental Settings

In the following sections, we evaluate the effectiveness of seven stopping criteria for active learning for WSD and TC tasks, including *max-confidence (MC)*, *min-error (ME)*, *overall-uncertainty (OU)*, *classification-change (CC)*, *strategy 1 (CC-MC)*, *strategy 2 (CC-ME)*, and *strategy 3 (CC-OU)*. Following previous studies (Zhu and Hovy, 2007), the predefined thresholds[3] used for MC, ME and OU are set to 0.01, 0.9 and 0.01, respectively.

To evaluate the effectiveness of each stopping criterion, we first construct two types of *baseline* methods called "*All*" and "*First*" methods. "All" method is defined as when all unlabeled examples in the pool are learned. "First" method is defined as when the current classifier reaches the same performance of the "All" method at the first time during the active learning process.

A better stopping criterion can not only achieve almost the same performance given by the "All" baseline method (i.e. *accuracy performance*), but also learn almost the same number of unlabeled examples by the "First" baseline method (i.e. *percentage performance*).

In uncertainty sampling scheme, the well-known entropy-based uncertainty measurement (Chen *et al.,* 2006; Schein and Ungar, 2007) is used in our active learning study as follows:

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

where $P(y|x)$ is the *a posteriori* probability. We denote the output class $y \in Y = \{y_1, y_2, \ldots, y_k\}$. *UM* is the uncertainty measurement function based on the entropy estimation of the classifier's posterior distribution.

We utilize maximum entropy (MaxEnt) model (Berger *et al.,* 1996) to design the basic classifier used in active learning for WSD and TC tasks. The advantage of the MaxEnt model is the ability to freely incorporate features from diverse sources into a single, well-grounded statistical

[3] In the following experiments, these thresholds are also used as initial values of β for individual criteria in the general combination strategy shown in Fig. 3.

model. A publicly available MaxEnt toolkit[4] was used in our experiments. To build the MaxEnt-based classifier for WSD, 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*, which are the same as the knowledge sources used in (Lee and Ng, 2002). In the design of text classifier, the maximum entropy model is also utilized, and no feature selection technique is used.

In the following active learning comparison experiments, the algorithm starts with a randomly chosen initial training set of 10 labeled examples, and makes 10 queries after each 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. In the following comparison experiments, the performance reported on Ontonotes data set is the macro-average on ten nouns, and the performance on TWA data set is the macro-average on six words.

5.2 Data Sets

Six publicly available natural data sets have been used in the following active learning comparison experiments. Three data sets are used for TC tasks: *WebKB*, *Comp2a* and *Comp2b*. The other three data sets are used for WSD tasks: *OntoNotes*, *Interest* and *TWA*.

The WebKB dataset was widely used in TC research. Following previous studies (McCallum and Nigam, 1998b), we use the four most populous categories: *student, faculty, course* and *project*, altogether containing 4199 web pages. In the preprocessing step, we only remove those words that occur merely once without using stemming. The resulting vocabulary has 23803 words.

The Comp2a data set consists of *comp.os.ms-windows.misc* and *comp.sys.ibm.pc.hardware* subset of NewsGroups. The Comp2b data set consists of *comp.graphics* and *comp.windows.x* categories from NewsGroups. Both two data sets have been previously used in active learning for TC (Roy and McCallum, 2001; Schein and Ungar, 2007).

The OntoNotes project (Hovy *et al*., 2006) uses the WSJ part of the Penn Treebank. The senses of noun words occurring in OntoNotes are linked to the Omega ontology. Ontonotes has been used previously in active learning for WSD tasks (Zhu and Hovy, 2007). In the following comparison experiments, we focus on 10 most frequent nouns[5] previously used in (Zhu and Hovy, 2007): *rate*, *president*, *people*, *part*, *point*, *director*, *revenue*, *bill*, *future*, and *order*.

The Interest data set developed by Bruce and Wiebe (1994) has been previously used for WSD (Ng and Lee, 1996). This data set consists of 2369 sentences of the noun "interest" with its correct sense manually labeled. The noun "*interest*" has six different senses in this data set. TWA developed by Mihalcea and Yang on 2003, is sense tagged data for six words with two-way ambiguities, previously used in WSD research. These six words are *bass*, *crane*, *motion*, *palm*, *plant* and *tank*. All instances were drawn from the British National Corpus.

5.3 Stopping Criteria for Uncertainty Sampling

In order to evaluate the effectiveness of our stopping criteria, we first apply them to uncertainty sampling for active learning for WSD and TC tasks. Table 1 shows that "First" method generally achieves higher performance than that of the "All" method. We can see from the "*Average*" row that stopping criteria MC, ME, CC-MC, CC-ME and CC-OU achieve close average accuracy performance to the "All" method whereas OU and CC achieve lower average accuracy performance. OU method achieves the lowest average accuracy performance. CC-ME achieves the highest average accuracy of 89.6%, followed by CC-MC.

Compared to the "First" method, CC-OU achieves the best average percentage performance of 37.03% (i.e. the closest one to the "First" method), followed by ME method. On six evaluation data sets, Table 1 shows that CC-ME method achieves 4 out of 6 highest accuracy performances, followed by CC-MC and MC methods. And CC-ME method also achieves 3 out of 6 best percentage performance, followed by CC, CC-OU and ME methods.

Among these four individual stopping criteria, ME outperforms MC, OU and CC. However, ME method can only be applied to batch-based selection because ME criterion is based on the feedback from Oracle. Too few informative candidates may not be adequate for obtaining a reasonable feedback for ME criterion.

[4]See http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

[5] See http://www.nlplab.com/ontonotes-10-nouns.rar

Data set	All	First	MC	ME	OU	CC	CC-MC	CC-ME	CC-OU
WebKB	0.910	0.911	0.910	0.910	0.837	0.912	0.912	**0.913**	0.912
	100%	31.50%	27.11%	29.11%	8.42%	**31.53%**	32.37%	33.02%	**31.53%**
Comp2a	0.880	0.884	0.877	0.879	0.868	0.876	0.879	**0.880**	0.876
	100%	35.12%	31.35%	31.28%	23.29%	27.35%	32.36%	**36.80%**	27.35%
Comp2b	0.900	0.901	0.887	0.888	0.880	0.879	0.891	**0.893**	0.882
	100%	41.66%	37.52%	36.76%	28.36%	30.80%	37.95%	**40.03%**	31.81%
Ontonotes	0.939	0.942	0.929	0.934	0.928	0.936	**0.940**	0.939	0.939
	100%	22.81%	30.19%	**22.14%**	21.81%	18.96%	34.77%	25.60%	24.75%
Interest	0.908	0.910	**0.910**	0.906	0.906	0.901	**0.910**	0.906	0.906
	100%	29.83%	37.54%	28.25%	28.51%	25.55%	37.54%	**28.67%**	28.62%
TWA	0.846	0.858	0.843	0.844	0.837	0.820	0.841	**0.845**	0.838
	100%	59.67%	80.34%	72.71%	70.47%	**61.54%**	86.99%	80.15%	78.12%
Average	0.897	0.901	0.892	0.893	0.876	0.887	0.895	**0.896**	0.892
	100%	37.43%	40.67%	36.71%	30.14%	32.62%	43.66%	40.71%	**37.03%**

Table 1. Effectiveness of seven stopping criteria for uncertainty sampling for active learning. For each data set, Table 1 shows the accuracy of the classifier and percentage of learned instances over all unlabeled data when each stopping criterion is met. The boldface numbers indicate the best corresponding performances.

Data set	All	MC	ME	OU	CC	CC-MC	CC-ME	CC-OU
WebKB	0.858	0.808	0.818	0.601	0.820	0.820	**0.824**	0.820
	100%	27.11%	29.11%	8.42%	31.53%	32.37%	33.02%	31.53%
Comp2a	0.894	0.838	0.839	0.825	0.837	0.838	**0.846**	0.837
	100%	31.35%	31.28%	23.29%	27.35%	32.36%	36.80%	27.35%
Comp2b	0.922	**0.884**	0.882	0.878	0.874	0.885	0.883	0.879
	100%	37.52%	36.76%	28.36%	30.80%	37.95%	40.03%	31.81%
Ontonotes	0.925	0.923	0.924	0.921	0.921	**0.932**	0.927	0.929
	100%	30.19%	22.14%	21.81%	18.96%	34.77%	25.60%	24.75%
Interest	0.899	**0.906**	0.890	0.890	0.885	**0.906**	0.891	0.890
	100%	37.54%	28.25%	28.51%	25.55%	37.54%	28.67%	28.62%
TWA	0.812	0.784	0.793	0.765	0.775	0.799	**0.810**	0.794
	100%	80.34%	72.71%	70.47%	61.54%	86.99%	80.15%	78.12%
Average	0.885	0.857	0.857	0.813	0.852	**0.863**	**0.863**	0.858
	100%	40.67%	36.71%	30.14%	32.62%	43.66%	40.71%	37.03%

Table 2. Effectiveness of seven stopping criteria for heterogeneous uncertainty sampling for active learning. Table 2 shows the accuracy of the classifier and percentage of learned instances over all unlabeled data when each stopping criterion is met. The boldface numbers indicate the best corresponding performances.

Interestingly, our proposed CC method acheves the best macro-average percentage performance on the TWA data set, however, other criteria work poorly, compared to the "First" method. Actually the sense distribution of each noun in TWA set is very skewed. From WSD experimental results on TWA, we found that only few learned instances can train the MaxEnt-based classifier with the highest accuracy performance.

In Table 1, the boldface numbers indicate the best performances. Three combination strategies achieve 12 out of 16 best performances[6]. We think the general combination strategy outperform individual stopping criteria for uncertainty sampling for active learning, because four individual stopping criteria only totally achieve 4 out of 16 best performances.

5.4 Stopping Criteria for Heterogeneous Uncertainty Sampling

In the following comparison experiments on heterogeneous uncertainty sampling, a MaxEnt-based classifier is used to select the most informative examples for training an another type of classifier based on multinomial naïve Bayes (NB) model (McCallum and Nigam, 1998b).

[6] CC and CC-OU methods achieve the same best percentage performance of 31.53% on WebKB data set. MC and CC-MC methods achieve the same highest accuracy performance of 91% on Interest data set.

Table 2 shows that the NB-based classifier trained on all data (i.e. "All method") achieves only 1.2% lower average accuracy performance than that of MaxEnt-based classifier. However, we can see from Table 2 that accuracy performances of each stopping criterion for heterogeneous uncertainty sampling are apparently lower than that for uncertainty sampling shown in Table 1. The main reason is that an uncertain example for one classifier (i.e. MaxEnt) may not be an uncertain example for other classifiers (i.e. NB). This comparison experiments aim to analyze the accuracy effectiveness of stopping criteria for heterogeneous uncertainty sampling, compared to that for uncertainty sampling shown in Table 1. Therefore we do not provide the results of the "First" method for heterogeneous uncertainty sampling. The "*Average*" row shows that CC-MC and CC-ME achieve the highest average accuracy performance of 86.3%, followed by CC-OU. On six data sets, CC-ME achieves 3 out of 6 highest accuracy performances.

Interestingly, these stopping criteria work very well on the Ontonotes and Interest data sets. Three combination strategies achieve higher accuracy performance than the "All" method on Ontonotes. However, the accuracy performances of these seven stopping criteria for heterogeneous uncertainty sampling on WebKB, Comp2a, Comp2b, and TWA degrade, compared to the "All" method.

The general combination strategy achieves 7 out of 9 boldface accuracy performances[7]. And only MC method achieves other 2 boldface accuracy performances. Experimental results show that the general combination strategy outperforms individual stopping criteria in overall for heterogeneous uncertainty sampling.

6 Related Work

Zhu and Hovy (2007) proposed a confidence-based framework to predict the upper bound and the lower bound for a stopping criterion in active learning. Actually this framework is a very coarse solution that simply uses *max-confidence* method to predict the upper bound, and uses *min-error* method to predict the lower bound. Zhu *et. al.* (2008) proposed a minimum expected error strategy to learn a stopping criterion through estimation of the classifier's expected error on future unlabeled examples. However, both two studies did not give an answer to the problem of how to define an appropriate threshold for the stopping criterion in a specific task.

Vlachos (2008) also studied a stopping criterion of active learning based on the estimate of the classifier's confidence, in which a separate and large dataset is prepared in advance to estimate the classifier's confidence. However, there is a risk to be misleading because how many examples are required for this pregiven separate dataset is an open question in real-world applications, and it can not guarantee that the classifier shows a rise-peak-drop confidence pattern during active learning process.

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 assumption seems to be unreasonable and difficult to satisfy. And their stopping criterion cannot be applied for active learning with other type of classifier such as NB, MaxEnt models.

7 Discussion

We believe that a classifier's performance change is a good signal of stopping the active learning process. It is worth studying further how to combine the factor of performance change with our proposed stopping criteria.

Among these stopping criteria, ME, CC, CC-ME can be used directly for committee-based sampling (Engelson and Dagan, 1999) for active learning. However, to use MC, OU, CC-MC and CC-OU for committee-based sampling, we should adopt a new uncertainty measurement such as *vote entropy* to measure the uncertainty of each unlabled example in the pool.

In the above active learning comparison experiments, the confidence estimation for each confidence-based stopping criterion is done within the unlabeled pool *U*. We think that for these confidence-based stopping criteria except SA method, confidence estimation on a large-scale outside unlabeled data set is worth studying in the future work.

8 Conclusion and Future Work

In this paper, we address the stopping criterion issue of active learning, and propose a new stopping criterion, classification-change, which considers the potential ability of each unlabeled ex-

[7] MC and CC-MC methods achieve the same highest accuracy performance of 90.6% on Interest data set. CC-MC and CC-CA methods achieve the same highest average accuracy performance of 86.3%.

ample on changing decision boundaries. To solve the problem of predefining an appropriate threshold for each confidence-based stopping criterion, a multi-criteria-based general combination strategy is proposed. Experimental results on uncertainty sampling and heterogeneous uncertainty sampling show that these stopping criteria work well on evaluation data sets, and combination strategies can achieve better performance than individual criteria. 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

This work was supported in part by the National 863 High-tech Project (2006AA01Z154) and the Program for New Century Excellent Talents in University (NCET-05-0287).

References

Berger Adam L., Vincent J. Della Pietra, Stephen A. Della Pietra. 1996. *A maximum entropy approach to natural language processing*. Computational Linguistics 22(1):39–71.

Bruce Rebecca and Janyce Wiebe. 1994. *Word sense disambiguation using decomposable models*. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 139-146.

Chen Jinying, Andrew Schein, Lyle Ungar and Martha Palmer. 2006. *An empirical study of the behavior of active learning for word sense disambiguation*. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 120-127

Engelson S. Argamon and I. Dagan. 1999. *Committee-based sample selection for probabilistic classifiers*. Journal of Artificial Intelligence Research (11):335-360.

Hovy Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. *Ontonotes: The 90% Solution*. In Proceedings of the Human Language Technology Conference of the NAACL, pp. 57-60.

Lee Yoong Keok and Hwee Tou Ng. 2002. *An empirical evaluation of knowledge sources and learning algorithm for word sense disambiguation*. In Proceedings of the ACL conference on Empirical methods in natural language processing, pp. 41-48

Lewis David D. and Jason Catlett. 1994. *Heterogeneous uncertainty sampling for supervised learning*. In Proceedings of 11th International Conference on Machine Learning, pp. 148-156

Lewis David D. and William A. Gale. 1994. *A sequential algorithm for training text classifiers*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 3-12

McCallum Andrew and Kamal Nigam. 1998a. *Employing EM in pool-based active learning for text classification*. In Proceedings of the 15th International Conference on Machine Learning, pp.350-358

McCallum Andrew and Kamal Nigam. 1998b. *A comparison of event models for naïve bayes text classification*. In AAAI-98 workshop on learning for text categorization.

Ng Hwee Tou and Hian Beng Lee. 1996. *Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp.40-47

Ngai Grace and David Yarowsky. 2000. *Rule writing or annotation: cost-efficient resource usage for based noun phrase chunking*. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 117-125

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

Schein Andrew I. and Lyle H. Ungar. 2007. *Active learning for logistic regression: an evaluation*. Machine Learning 68(3): 235-265

Schohn Greg and David Cohn. 2000. *Less is more: Active learning with support vector machines*. In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 839-846

Shen Dan, Jie Zhang, Jian Su, Guodong Zhou and Chew-Lim Tan. 2004. *Multi-criteria-based active learning for named entity recognition*. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

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

Vlachos Andreas. 2008. *A stopping criterion for active learning*. Computer Speech and Language. 22(3): 295-312

Zhu Jingbo and Eduard Hovy. 2007. *Active learning for word sense disambiguation with methods for addressing the class imbalance problem*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 783-790

Zhu Jingbo, Huizhen Wang and Eduard Hovy. 2008. *Learning a stopping criterion for active learning for word sense disambiguation and text classification*. In Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 366-372