Active Learning with Sampling by Uncertainty and Density for Word Sense Disambiguation and Text Classification

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

Benjamin K Tsou
Language Information Sciences
Research Centre
City University of Hong Kong
HK, P.R.China
rlbtsou@cityu.edu.hk

Abstract

This paper addresses two issues of active learning. Firstly, to solve a problem of uncertainty sampling that it often fails by selecting outliers, this paper presents a new selective sampling technique, sampling by uncertainty and density (SUD), in which a k-Nearest-Neighbor-based density measure is adopted to determine whether an unlabeled example is an outlier. Secondly, a technique of sampling by clustering (SBC) is applied to build a representative initial training data set for active learning. Finally, we implement a new algorithm of active learning with SUD and SBC techniques. The experimental results from three real-world data sets show that our method outperforms competing methods, particularly at the early stages of active learning.

1 Introduction

Creating a large labeled training corpus is expensive and time-consuming in some real-world applications (e.g. word sense annotation), and is often a bottleneck to build a supervised classifier for a new application or domain. Our study aims to minimize the amount of human labeling efforts required for a supervised classifier (e.g. for automated word sense disambiguation) to achieve a satisfactory performance by using *active learning*.

Among the techniques to solve the knowledge bottleneck problem, active learning is a widely used framework in which the learner has the ability to automatically select the most informative unlabeled examples for human annotation. The ability of the active learner can be referred to as *selective sampling*. *Uncertainty sampling* (Lewis and Gale, 1994) is a popular selective sampling technique, and has been widely studied in natural language processing (NLP) applications such as word sense disambiguation (WSD) (Chen *et al.*, 2006; Chan and Ng, 2007), text classification (TC) (Lewis and Gale, 1994; Zhu *et al.*, 2008), statistical syntactic parsing (Tang *et al.*, 2002), and named entity recognition (Shen *et al.*, 2004).

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. However, uncertainly sampling often fails by selecting outliers (Roy and McCallum, 2001; Tang *et al.*, 2002). These selected outliers (i.e. unlabeled examples) have high uncertainty, but can not provide much help to the learner. To solve the outlier problem, we proposed in this paper a new method, *sampling by uncertainty and density* (SUD), in which a *K-Nearest-Neighbor-based density* (KNN-density) measure is used to determine whether an unlabeled example is an outlier, and a combination strategy based on KNN-density measure and uncertainty measure is designed to select the most informative unlabeled examples for human annotation at each learning iteration.

The second effort we made is to study how to build a representative initial training data set for active learning. We think building a more representative initial training data set is very helpful for active learning. In previous studies on active learning, the initial training data set is generally generated at random, based on an assumption that random sampling will be likely to build the initial training set with same prior data distribution as that of whole corpus. However, this situation seldom occurs in real-world applications due to the small size of initial training set used. In

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1137–1144
Manchester, August 2008

this paper, we utilize an approach, *sampling by clustering* (SBC), to selecting the most representative examples to form initial training data set for active learning. To do it, the whole unlabeled corpus should be first clustered into predefined number of clusters (i.e. the predefined size of the initial training data set). The example closest to the centroid of each cluster will be selected to augment initial training data set, which is considered as the most representative case.

Finally, we describe an implementation of active learning with SUD and SBC techniques. Experimental results of active learning for WSD and TC tasks show that our proposed method outperforms competing methods, particularly at the early stages of active learning process. It is noteworthy that these proposed techniques are easy to implement, and can be easily applied to several learners, such as Maximum Entropy (ME), naïve Bayes (NB) and Support Vector Machines (SVMs).

2 Active Learning Process

In this work, we are interested in *uncertainty sampling* (Lewis and Gale, 1994) for pool-based active learning, in which an unlabeled example x with maximum uncertainty is selected 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.

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[2] 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

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

3 Uncertainty Measures

In real-world applications, only limited size of training sample set can be provided to train a supervised classifier. Due to manual efforts involved, such brings up a considerable issue: what is the best subset of examples to annotate. In the uncertainty sampling scheme, the unlabeled example with maximum uncertainty is viewed as the most informative case. The key point of uncertainty sampling is how to measure the uncertainty of an unlabeled example x.

3.1 Entropy Measure

The well-known *entropy* is a popular uncertainty measurement widely used in previous studies on active learning (Tang *et al.*, 2002; Chen *et al.* 2006; Zhu and Hovy, 2007):

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

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

In the following comparison experiments, the uncertainty sampling based on entropy criterion is considered as the baseline method, also called traditional uncertainty sampling.

3.2 Density*Entropy Measure

To analyze the outlier problem of traditional uncertainty sampling, we first give an example to explain our motivation.

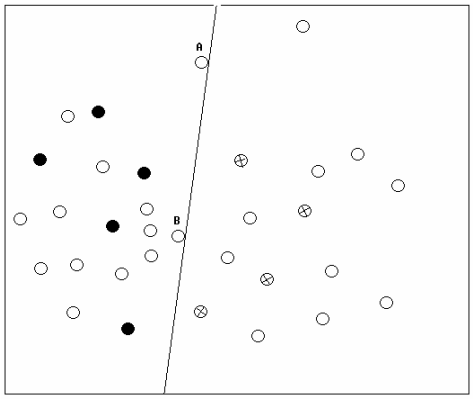

Figure 2. An example of two points A and B with maximum uncertainty at the i[th] learning iteration

As mentioned in Section 1, the motivation behind uncertainty sampling is to find some unlabeled examples near decision boundaries, and assume that these examples have the maximum uncertainty. Fig. 2 shows two unlabeled examples A and B with maximum uncertainty at the i[th]

learning cycle. Roughly speaking, there are three unlabeled examples near or similar to *B*, but, none for *A*. We think example B has higher representativeness than example A, and A is likely to be an outlier. We think adding B to the training set will help the learner more than A.

The motivation of our study is that we prefer not only the most informative example in terms of uncertainty measure, but also the most representative example in terms of density measure. The density measure can be evaluated based on how many examples there are similar or near to it. An example with high density degree is less likely to be an outlier.

In most real-world applications, because the scale of unlabeled corpus would be very large, Tang *et al.* (2002) and Shen *et al.* (2004) evaluated the density of an example within a cluster. Unlike their work[3], we adopt a new approach, called *K-Nearest-Neighbor-based density* (KNN-density) measure, to evaluating the density of an unlabeled example *x*. Given a set of *K* (i.e. =20 used in our experiments) most similar examples $S(x)=\{s_1, s_2, \ldots, s_K\}$ of the example *x*, the KNN-density DS(.) of example *x* is defined as:

$$DS(x) = \frac{\sum_{s_i \in S(x)} \cos(x, s_i)}{K} \quad (2)$$

As discussed above, we prefer to select examples with maximum uncertainty and highest density for human annotation. We think getting their labels can help the learner greatly. To do it, we proposed a new method, *sampling by uncertainty and density* (SUD), in which entropy-based uncertainty measure and KNN-density measure are considered simultaneously.

In SUD scheme, a new uncertainty measure, called *density*entropy* measure[4], is defined as:

$$DSH(x) = DS(x) \times H(x) \quad (3)$$

4 Initial Training Set Generation

As shown in Fig. 1, only a small number of training samples are provided at the beginning of active learning process. In previous studies on active learning, the initial training set is generally generated by random sampling from the whole unlabeled corpus. However, random sampling technique can not guarantee selecting a most representative subset, because the size of initial training set is generally too small (e.g. 10). We think selecting some representative examples to form initial training set can help the active learner.

In this section we utilize an approach, *sampling by clustering* (SBC), to selecting the most representative examples to form initial training data set. In the SBC scheme, the whole unlabeled corpus has been first clustered into a predefined number of clusters (i.e. the predefined size of the initial training set). The example closest to the centroid of each cluster will be selected to augment initial training set, which is viewed as the most representative case.

We use the K-means clustering algorithm (Duda and Hart, 1973) to cluster examples in the whole unlabeled corpus. In the following K-means clustering algorithm, the traditional cosine measure is adopted to estimate the similarity between two examples, that is

$$\cos(w_i, w_j) = \frac{w_i \bullet w_j}{\|w_i\| \cdot \|w_j\|} \quad (4)$$

where w_i and w_j are the feature vectors of the examples *i* and *j*.

To summarize the SBC-based initial training set generation algorithm, let $U=\{U_1, U_2, \ldots, U_N\}$ be the set of unlabeled examples to be clustered, and *k* be the predefined size of initial training data set. In other words, SBC technique selects *k* most representative unlabeled examples from *U* to generate the initial training data set. The SBC-based initial training set generation procedure is summarized as follows:

SBC-based Initial Training Set Generation

Input: *U*, *k*

Phrase 1: Cluster the corpus *U* into *k* clusters Ψ_j(j=1,…,*k*) by using K-means clustering algorithm as follows:

1. Initialization. Randomly choosing *k* examples as the *centroid* ϕ_j(j=1,…,*k*) for initial clusters Ψ_j(j=1,…,*k*), respectively.
2. Re-partition $\{U_1, U_2, \ldots, U_N\}$ into *k* clusters Ψ_j(j=1,…,*k*), where
   $\Psi_j = \{U_i : \cos(U_i, \phi_j) \geq \cos(U_i, \phi_t), t \neq j\}$.
3. Re-estimate the *centroid* ϕ_j for each clusters Ψ_j, that is:
   $\phi_j = \frac{\sum_{U_i \in \Psi_j} U_i}{m}$, where *m* is the size of Ψ_j.
4. Repeat Step 2 and Step 3 until the algorithm converges.

[3] We also tried their cluster-based density measure, but performance was essentially degraded.

[4] We also tried other ways like λ **DS(x)+(1-* λ *) H(x)* measure used in previous studies, but it seems to be random. Actually it is very difficult to determine an appropriate λ value for a specific task.

Phrase 2: Select the example u_j closest to the centroid ϕ_j for each cluster Ψ_j to augment initial training data set Ω, where

$\Omega = \{u_j : \cos(u_j, \phi_j) \geq \cos(U_i, \phi_j), u_j \neq U_i, j \in [1, k]\}$

Return Ω;

The computation complexity of the K-means clustering algorithm is *O(NdkT)*, where *d* is the number of features and *T* is the number of iterations. In practice, we can define the stopping criterion (i.e. shown in Step 4) of K-means clustering algorithm that relative change of the total distortion is smaller than a threshold.

5 Active Learning with SUD and SBC

Procedure: Active Learning with SUD and SBC
Input: Pool of unlabeled data set *U*; *k* is the predefined size of initial training data set
Initialization.

- Evaluate the density of each unlabeled example in terms of *KNN-density* measure;
- Use *SBC* technique to generate the small initial training data set of size *k*.

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 in terms of *density*entropy* measure to select *m* most informative unlabeled examples, and ask oracle *H* for labeling, namely SUD scheme.
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 3. Active learning with SUD and SBC

Fig. 3 shows the algorithm of active learning with SUD and SBC techniques. Actually there are some variations. For example, if the initial training data set is generated by SBC, and entropy-based uncertainty measure is used, it is active learning with SBC. Similarly, if the initial training data set is generated at random, and the density*entropy uncertainty measure is used, it is active learning with SUD. If both SBC and SUD techniques are not used, we call it (traditional) uncertainty sampling as baseline method.

6 Evaluation

In the following comparison experiments, we evaluate the effectiveness of various active learning methods for WSD and TC tasks on three publicly available real-world data sets.

6.1 Deficiency Measure

To compare various active learning methods, *deficiency* is a statistic developed to compare performance of active learning methods globally across the learning curve, which has been used in previous studies (Schein and Unga, 2007). The deficiency measure can be defined as:

$$Def_n(AL, REF) = \frac{\sum_{t=1}^{n}(acc_n(REF) - acc_t(AL))}{\sum_{t=1}^{n}(acc_n(REF) - acc_t(REF))} \quad (5)$$

where acc_t is the average accuracy at t^{th} learning iteration. REF is the baseline active learning method, and AL is the active learning variant of the learning algorithm of REF, e.g. active learning with SUD and SBC. *n* refers to the evaluation stopping points (i.e. the number of learned examples). Smaller deficiency value (i.e. <1.0) indicates AL method is better than REF method. Conversely, a larger value (i.e. >1.0) indicates a negative result.

In the following comparison experiments, we evaluate the effectiveness of six active learning methods, including *random sampling (random)*, *uncertainty sampling (uncertainty)*, *SUD*, *random sampling with SBC (random+SBC)*, *uncertainty sampling with SBC (uncertainty+SBC)*, and *SUD with SBC (SUD+SBC)*. "+SBC" indicates initial training data set generated by SBC technique. Otherwise, initial training set is generated at random. To evaluate deficiency of each method, the REF method (i.e. the baseline method) defined in Equation (5) refers to (traditional) uncertainty sampling.

6.2 Experimental Settings

We utilize a maximum entropy (ME) model (Berger *et al.*, 1996) to design the basic classifier for WSD and TC tasks. The advantage of the ME model is the ability to freely incorporate features from diverse sources into a single, well-grounded statistical model. A publicly available ME toolkit[5] was used in our experiments. To build the ME-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.

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

In the following comparison experiments, the algorithm starts with a initial training set of 10 labeled examples, and make 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.

6.3 Data Sets

Three publicly available natural data sets have been used in the following active learning comparison experiments. *Interest* data set is used for WSD tasks. *Comp2* and *WebKB* data sets are used for TC tasks.

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.

The Comp2 data set consists of *comp.graphics* and *comp.windows.x* categories from News-Groups, which has been previously used in active learning for TC (Roy and McCallum, 2001; Schein and Ungar, 2007).

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

Data sets	Interest	Comp2	WebKB
Accuracy	0.908	0.90	0.91

Table 1. Average accuracy of supervised learning on each data set when all examples have been learned.

6.4 Active Learning for WSD Task

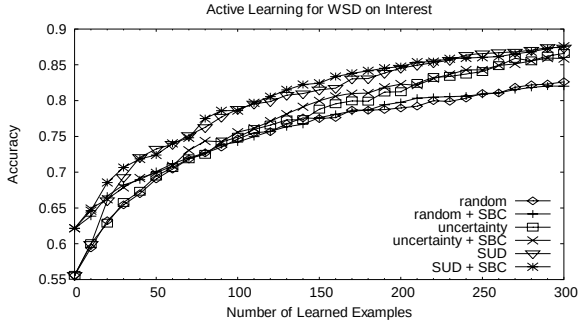

Figure 4. Active learning curve for WSD on Interest data set

Random	Random+SBC	Uncertainty
1.926	1.886	NA
Uncertainty+SBC	SUD	SUD+SBC
0.947	0.811	**0.758**

Table 2. Average *deficiency* achieved by various active learning methods on Interest data set. The stopping point is 300.

Fig. 4 depicts performance curves of various active learning methods for WSD task on Interest data set. Among these six methods, random sampling method shows the worst performance. SUD method constantly outperforms uncertainty sampling. As discussed above, SUD method prefers not only the most uncertainty examples, but also the most representative examples. In the SUD scheme, the factor of KNN-density can effectively avoid selecting the outliers that often cause uncertainty sampling to fail.

It is noteworthy that using SBC to generate initial training data set can improve random (-0.04 deficiency), uncertainty (-0.053 deficiency) and SUD (-0.053 deficiency) methods, respectively. If the initial training data set is generated at random, the initial accuracy is only 55.6%. Interestingly, SBC achieves 62.2% initial accuracy, and makes 6.6% accuracy performance improvement. However, SBC only makes performance improvement for each method at the early stages of active learning. After 50 unlabeled examples have been learned, it seems that SBC has very little contribution to random, uncertainty and SUD methods. Table 2 shows that the best method is SUD with SBC (0.758 deficiency), followed by SUD method.

6.5 Active Learning for TC Tasks

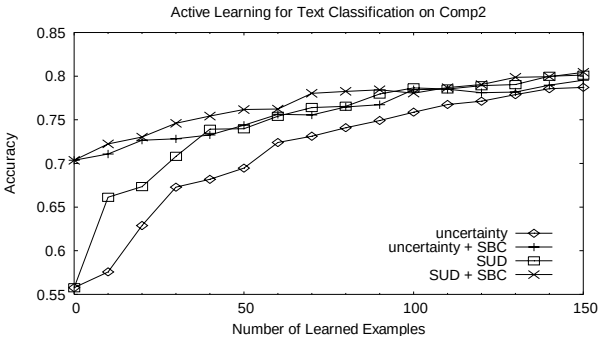

Figure 5. Active learning curve for text classification on Comp2 data set

Uncertainty	Uncertainty+SBC	SUD	SUD+SBC
NA	0.409	0.588	**0.257**

Table 3. Average *deficiency* achieved by various active learning methods on Comp2 data set. The stopping point is 150.

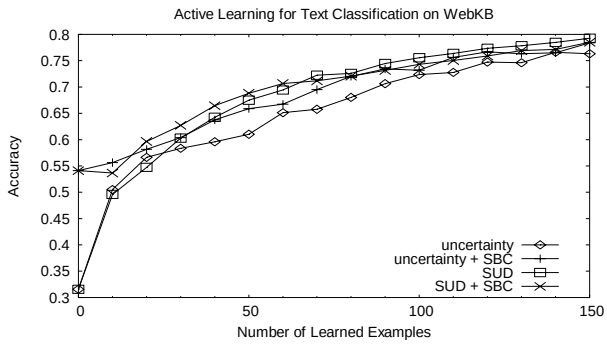

Figure 6. Active learning curve for text classification on WebKB data set

Uncertainty	Uncertainty+SBC	SUD	SUD+SBC
NA	0.669	0.748	**0.595**

Table 4. Average *deficiency* achieved by various active learning methods on WebKB data set. The stopping point is 150.

Fig. 5 and 6 show the effectiveness of various active learning methods for text classification tasks. Since random sampling performs poorly as shown in Fig. 4, it is not further shown in Fig. 5 and 6. We only compare uncertainty sampling and our proposed methods for both text classification tasks.

Similarly, SUD method constantly outperforms uncertainty sampling on two data sets. SBC greatly improves uncertainty sampling (i.e. 0.591 and 0.331 deficiencies degraded) and SUD method (i.e. 0.331 and 0.153 deficiencies degraded), respectively. Interestingly, unlike WSD task shown in Fig. 4, Table 3 and 4 show that uncertainty sampling with SBC outperforms our SUD method for text classification on both data sets. The reason is that SBC makes about 15% initial accuracy improvement on Comp2 data set, and about 23% initial accuracy improvement on WebKB data set. Such improvements indicate that selecting high representative initial training set is very necessary and helpful for active learning. Table 3 and 4 show that the best active learning method for TC task is SUD with SBC, following by uncertainty sampling with SBC method. It is noteworthy that on WebKB uncertainty sampling with SBC (0.669 deficiency) achieves only slight better performance than SUD method (0.748 deficiency) as shown in Table 4, simply because SBC only introduce good performance improvement at the early stages. Actually on WebKB SUD method achieves slight better performance than uncertainty sampling with SBC after about 50 unlabeled examples have been learned.

7 Related Work

In recent years active learning has been widely studied in various natural language processing (NLP) tasks, such as word sense disambiguation (Chen *et al.*, 2006; Zhu and Hovy, 2007), text classification (TC) (Lewis and Gale, 1994; McCallum and Nigam, 1998), named entity recognition (NER) (Shen *et al.*, 2004), chunking (Ngai and Yarowsky, 2000), information extraction (IE) (Thompson *et al.*, 1999), and statistical parsing (Tang *et al.*, 2002).

In addition to uncertainty sampling, there is another popular selective sampling scheme, *Query-by-committee* (Engelson and Dagan, 1999), which generates a committee of classifiers (always more than two classifiers) and selects the next unlabeled example by the principle of maximal disagreement among these classifiers. A method similar to committee-based sampling is *co-testing* proposed by Muslea *et al.* (2000), which trains two learners individually on two compatible and uncorrelated views that should be able to reach the same classification accuracy. In practice, however, these conditions of view selection are difficult to meet in real-world applications. Cohn *et al.* (1996) and Roy and McCallum (2001) proposed a method that directly optimizes expected future error on future test examples. However, the computational complexity of their methods is very high.

There are some similar previous studies (Tang *et al.*, 2002; Shen *et al.*, 2004) in which the representativeness criterion in active learning is considered. Unlike our sampling by uncertainty and density technique, Tang *et al.* (2002) adopted a sampling scheme of *most uncertain per cluster* for NLP parsing, in which the learner selects the sentence with the highest uncertain score from each cluster, and use the density to weight the selected examples while we use density information to select the most informative examples. Actually the scheme of most uncertain per cluster still can not solve the outlier problem faced by uncertainty sampling technique. Shen *et al.* (2004) proposed an approach to selecting examples based on informativeness, representativeness and diversity criteria. In their work, the density of an example is evaluated within a cluster, and multiple criteria have been linearly combined with some coefficients. However, it is difficult to automatically determine sufficient coefficients in real-world applications. Perhaps there are different appropriate coefficients for various applications.

8 Discussion

For batch mode active learning, we found sometimes there is a redundancy problem that some selected examples are identical or similar. Such situation would reduce the representativeness of selected examples. To solve this problem, we tried the sampling scheme of "most uncertain per cluster" (Tang *et al.,* 2002) to select the most informative examples. We think selecting examples from each cluster can alleviate the redundancy problem. However, this sampling scheme works poorly for WSD and TC on the three data sets, compared to traditional uncertainty sampling. From the clustering results, we found these resulting clusters are very imbalanced. It makes sense that more informative examples are contained in a bigger cluster. In this work, we only use SUD technique to select the most informative examples for active learning. We plan to study how combining SBC and SUD techniques can enhance the selection of the most informative examples in the future work.

Furthermore, we think that a misclassified unlabeled example may convey more information than a correctly classified unlabeled example which is closer to the decision boundary. But there is a difficulty that the true label of each unlabeled example is unknown. To use misclassification information to select the most informative examples, we should study how to automatically determine whether an unlabeled example has been misclassified. For example, we can make an assumption that an unlabeled example may be misclassified if this example was previously "outside" and is now "inside". We will study this issue in the future work.

Actually these proposed techniques can be easily applied for committee-based sampling for active learning. However, to do so, we should adopt a new uncertainty measurement such as *vote entropy* to measure the uncertaity of each unlabled example in committee-based sampling scheme.

9 Conclusion and Future Work

In this paper, we have addressed two issues of active learning, involving the outlier problem of traditional uncertainty sampling, and initial training data set generation. To solve the outlier problem of traditional uncertainly sampling, we proposed a new method of sampling by uncertainty and density (SUD) in which KNN-density measure and uncertainty measure are combined together to select the most informative unlabeled example for human annotation at each learning cycle. We employ a method of sampling by clustering (SBC) to generate a representative initial training data set. Experimental results on three evaluation data sets show that our combined SUD with SBC method achieved the best performance compared to other competing methods, particularly at the early stages of active learning process. In future work, we will focus on the redundancy problem faced by batch mode active learning, and how to make use of misclassified information to select the most useful examples for human annotation.

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.

Chan Yee Seng and Hwee Tou Ng. 2007. *Domain adaptation with active learning for word sense disambiguation*. Proceedings of the 45th annual meeting on Association for Computational Linguistics, pp. 49-56

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

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

Duda Richard O. and Peter E. Hart. 1973. *Pattern classification and scene analysis*. New York: Wiley.

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

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-02 conference on Empirical methods in natural language processing, pp. 41-48

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. 1998. *A comparison of event models for naïve bayes text classification*. In AAAI-98 workshop on learning for text categorization.

Muslea Ion, Steven Minton and Craig A. Knoblock. 2000. *Selective sampling with redundant views*. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 621-626.

Ng Hwee Tou and Hian Beng Lee. 1996. *Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach*. In Proceedings of the Thirty-Fourth 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

Thompson Cynthia A., Mary Elaine Califf and Raymond J. Mooney. 1999. *Active learning for natural language parsing and information extraction*. In Proceedings of the Sixteenth International Conference on Machine Learning, pp. 406-414

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