Applying active learning to supervised word sense disambiguation in MEDLINE

Yukun Chen,[1] Hongxin Cao,[2] Qiaozhu Mei,[3,4] Kai Zheng,[4,5] Hua Xu[1,6]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001244).

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[2]Department of Medical Informatics, Second Military Medical University, Shanghai, China
[3]School of Information, University of Michigan, Ann Arbor, Michigan, USA
[4]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, USA
[5]Department of Health Management and Policy, University of Michigan, Ann Arbor, Michigan, USA
[6]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Correspondence to
Dr Hua Xu, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA; hua.xu@uth.tmc.edu

YC and HC contributed equally to this work.

Received 4 August 2012
Revised 17 October 2012
Accepted 6 January 2013
Published Online First
30 January 2013

To cite: Chen Y, Cao H, Mei Q, *et al. J Am Med Inform Assoc* 2013;**20**:1001–1006.

ABSTRACT

Objectives This study was to assess whether active learning strategies can be integrated with supervised word sense disambiguation (WSD) methods, thus reducing the number of annotated samples, while keeping or improving the quality of disambiguation models.

Methods We developed support vector machine (SVM) classifiers to disambiguate 197 ambiguous terms and abbreviations in the MSH WSD collection. Three different uncertainty sampling-based active learning algorithms were implemented with the SVM classifiers and were compared with a passive learner (PL) based on random sampling. For each ambiguous term and each learning algorithm, a learning curve that plots the accuracy computed from the test set as a function of the number of annotated samples used in the model was generated. The area under the learning curve (ALC) was used as the primary metric for evaluation.

Results Our experiments demonstrated that active learners (ALs) significantly outperformed the PL, showing better performance for 177 out of 197 (89.8%) WSD tasks. Further analysis showed that to achieve an average accuracy of 90%, the PL needed 38 annotated samples, while the ALs needed only 24, a 37% reduction in annotation effort. Moreover, we analyzed cases where active learning algorithms did not achieve superior performance and identified three causes: (1) poor models in the early learning stage; (2) easy WSD cases; and (3) difficult WSD cases, which provide useful insight for future improvements.

Conclusions This study demonstrated that integrating active learning strategies with supervised WSD methods could effectively reduce annotation cost and improve the disambiguation models.

INTRODUCTION

Word sense disambiguation (WSD) is the process of identifying the appropriate sense of an ambiguous word in a given context. WSD is important for many natural language processing (NLP) tasks, such as information extraction and information retrieval.[1] The ambiguity inherent in biomedical texts is a widely recognized problem. For example, 'gene,' an important entity in biomedical research, can have ambiguous names referring to: (1) multiple genes; (2) a gene or an English word not related to a gene; (3) RNA, protein, or a gene; or (4) genes in different species. A gene name ambiguity study showed that 85.1% of correctly retrieved mouse genes were ambiguous, easily confused with other gene names from 21 organisms in a set of 45 000 abstracts associated with mouse genes.[2]

Many different approaches have been developed for biomedical WSD tasks, as described in a review paper by Schuemie *et al.*[3] Among them, supervised machine learning-based WSD methods have received considerable attention and have shown very good results in both general English texts[4–8] and biomedical texts such as MEDLINE abstracts.[9] Supervised WSD approaches usually build a classification model for each ambiguous word by learning from an annotated corpus containing instances of each possible sense of the word. Despite its high performance, supervised WSD has limited scalability as it is a costly and time-consuming process to build a sense-annotated corpus for each ambiguous term in biomedical texts. Researchers have investigated different automated methods to create pseudo-corpora with labeled senses and have used them for supervised WSD methods (also called semi-supervised).[10 11] Despite the successes, WSD methods based on pseudo-corpora did not perform as well as supervised WSD systems that were based on annotated instances from the real corpus.[3] An alternative new approach presented in this study is to investigate how active learning strategies can be integrated with supervised WSD methods to reduce the number of annotated samples required by a satisfactory classification model.

Active learning, an approach that uses an active sampling algorithm, is one of the possible solutions in many supervised learning tasks where annotated samples are expensive to obtain. Applying active learning to classification could improve efficiency in constructing the classification model.[12] An active learner (AL), using a learning system with an active learning algorithm implemented, is capable of achieving better learning performance with less learning cost by actively selecting the queries (instances) for labeling rather than choosing them randomly. Researchers have applied active learning to many areas including text classification, information extraction, image classification and retrieval, etc.[13–16] More specifically, many NLP tasks that require large numbers of annotated samples have also benefited from active learning.[17–20] For WSD tasks in the general English domain, Zhu and Hovy[21] proposed an over-sampling method and it worked better than ordinary uncertainty sampling for WSD on 38 randomly chosen ambiguous nouns with imbalanced classes.

Few studies, however, have applied active learning to biomedical classification tasks. One of our recent studies investigated the application of active learning to the assertion classification of concepts in clinical texts. The result showed that the active learning strategy could achieve the same classification performance as the random sampling approach and use about 60% fewer annotated samples.[22] More recently, another study applied active learning to a few clinical text classification tasks and showed that it was better than the random sampling method.[23] To the best of our knowledge, no study has explored the use of active learning in supervised WSD tasks in the biomedical domain. It is not known whether active learning can be helpful to biomedical WSD tasks by reducing annotation costs and improving classification quality.

In this paper, we describe a study where we applied three different active learning algorithms to support vector machine (SVM)-based disambiguation models for 197 ambiguous terms from MEDLINE abstracts. We compared learning curves between three ALs and a passive learner (PL), based on random sampling across 197 WSD tasks. Our evaluation showed that all of the ALs were statistically significantly better than the PL, indicating that integrating active learning strategies with supervised WSD methods could effectively reduce annotation cost and improve the quality of disambiguation models.

METHODS

Three different uncertainty sampling-based active learning algorithms (Least Confidence (LC), Margin, and Entropy) and one passive learning method (random sampling) were integrated with an SVM classifier to disambiguate 197 ambiguous words and abbreviations in the MSH WSD collection derived from MEDLINE abstracts. For each ambiguous term and for each learning algorithm, an average learning curve was generated that plots the accuracy computed from the test set as a function of the number of annotated samples used in the model via a 10-fold cross-validation (CV). The area under the average learning curve was used as the primary metric for evaluation.

WSD dataset

In this study, we used the MSH WSD dataset developed by Jimeno-Yepes *et al*.[24] This benchmark dataset was downloaded from the National Library of Medicine (NLM) WSD test collection collaboration.[25] The generation of MSH WSD is based on exploiting MeSH indexing in MEDLINE abstracts. It consists of 106 ambiguous abbreviations, 88 ambiguous terms and 9 which are a combination of both, for a total of 203 ambiguous words.[24] Each instance containing the ambiguous word is assigned an appropriate sense that is represented using a concept unique identifier (CUI) from the 2009AB version of the UMLS (Unified Medical Language System). For each ambiguous term/abbreviation, the dataset contains a maximum of 100 instances obtained from MEDLINE for each sense, resulting in 37 888 ambiguous cases in 37 090 MEDLINE citations.[24] In the study by Jimeno-Yepes *et al*,[24] the authors also evaluated machine learning-based WSD algorithms on this dataset and reported an accuracy of 0.9386 for the entire MSH WSD dataset, when words from titles and abstracts were used as features. To ensure that we had enough samples for training and testing, we included ambiguous words that have more than 100 instances in total for all senses in this study, resulting in 197 words. Among them, 111 are abbreviations and 86 are unabbreviated terms. In addition, 14 out of the 197 words have more than two senses and the remaining 183 words have exactly two senses. Table A1 in the online appendix shows the frequency distribution of the senses for each ambiguous word in the dataset.

ACTIVE LEARNING-ENABLED SUPERVISED WSD

The pool-based active learning approach to classification

An active learning-based classification system mainly consists of two core components: a classification model and an active sample selection or a querying model. The pool-based active learning approach to classification[13] was used in this study. The approach starts with a pool of unlabeled samples and it iteratively selects informative samples for annotation and model development. The following process describes the pool-based active learning experiment for a given dataset and a querying algorithm:

1. Initialize the labeled training set $L=L_0$, the pool of unlabeled set $U=U_0$, and a test set T.
2. Train the classification model based on L and predict the class label for each instance in U and T.
3. Rank the instances in U based on the querying algorithm and assign labels (from human experts) for the top b(i) samples in U, where b(i), the batch size of active learning, is the number of querying samples at iteration i.
4. Add the b(i) instance(s) with label(s) to L and remove from U.
5. Compute the classification performance in accuracy (ACC) on the test set T and store in ACC(i).
6. Iterate steps (2) to (5) until the stop criterion (eg, unlabeled samples in the pool are used up) is met.
7. Evaluate this learning process by using the global learning score based on the learning curve that plots ACC(i) as a function of the batch size b(i). The 'Evaluation' section describes this evaluation metric in detail.

In the MSH WSD dataset, the pool size varies from 100 to 500, depending on the ambiguous word. We pretended that labels of samples were not available when running the querying algorithms. For the initial training set, we randomly selected two samples from the entire pool. All experiments with different querying algorithms used the same initial training set and, therefore, have the same initial point in the learning curve. In this study, we used a batch size of 1 in all experiments so that we could closely monitor the performance increase by every incremental training sample. As the minimum number of training samples for an ambiguous word was 100 and we used 10-fold CV in the evaluation (see the 'Evaluation' section), we stopped the active learning process when 90 training samples were queried.

The WSD classification model

The WSD classification model was built on the SVM algorithm with linear kernel in the package 'Liblinear.'[26] We used a one-vs.-all multi-class classification model if the ambiguous term has more than two senses. As optimized parameters of SVM classifiers were different for the 197 words in the dataset, we used a common setting: s=1 (L2-regularized L2-loss support vector classification) and c=1, for all words in this study, which performed comparably to the previous study.[24] The numeric outputs by SVM classifiers were mapped into the probabilistic domain (values from 0 to 1) by a sigmoid/logistic function. All words (except the ambiguous word itself) occurring in the title and abstract of a citation where the ambiguous term appears were used as features for SVM classifiers, similarly to the previously reported study.[24]

Active learner and passive learner

The second core component of active learning is the querying method. In general, there are two types of learners: AL and PL. The PL randomly queries instances from the pool of unlabeled samples, without considering the information about samples in the pool. The AL, on the other hand, will select the instances that are the most promising, improving the predictive performance of the model. x^* is selected as the most informative sample according to the function x^*=argmax Q(x), where Q(x) is the querying algorithm that outputs the informativeness or querying value (Q value) for data matrix x in U. In this study, we implemented three uncertainty sampling-based querying algorithms that query the sample with the least certainty or closest to the decision boundary. They are appropriate for multi-class classification problems such as supervised WSD tasks.

The simplest uncertainty sampling algorithm is called LC, which is straightforward for the probabilistic models:

$$Q^{LC}(x) = 1 - P(y^*|x;\theta)$$

where y* is the most likely label sequence for x and θ is the model that generates the posterior probability P of label y given data matrix x. In the binary classification case, LC is equivalent to querying the instance with the highest Q value (or uncertainty value) that is nearest the 0.5 posterior probability of being in the positive or negative class.

As LC only considers information about the most probable label, we also used a different multi-class uncertainty sampling method called margin sampling (Margin):

$$Q^{margin}(x) = P(y_2^*|x;\theta) - P(y_1^*|x;\theta),$$

where y_1* and y_2* are the first and second most probable class labels under the model, respectively. The intuition of this algorithm is that the samples with larger margins are easier to differentiate between the two most likely class labels, while the samples with smaller margins are more ambiguous. Thus, the margin sampling algorithm outputs the sample with the smallest difference between the two most likely class labels.

For problems with very large label sets, however, the margin method still ignores much of the output distribution for the remaining classes. Thus we implemented a more general uncertainty sampling strategy called Entropy:

$$Q^{entropy}(x) = -\sum_i P(y_i|x;\theta) \log P(y_i|x;\theta)$$

where y_i ranges over all possible labels. Entropy is a measure of uncertainty or impurity over all possible labels in a machine-learning task.

For binary classification, all three are equivalent to querying the instance with a class posterior closest to 0.5. All three querying algorithms were expected to have identical performance on 183 ambiguous words that have only two possible senses. Therefore, we focused the comparison study among three querying algorithms only on the 14 ambiguous words with more than two senses.

Evaluation

In this study, we used evaluation measurements similar to those in the 2010 active learning challenge.[27] The performance of the active learning-enabled classification system was evaluated by a learning curve, which plotted the accuracy (ACC) computed using the test set as a function of the number of labels annotated. ACC was defined as the ratio between the number of correctly identified samples and the number of all samples in the test set. A commonly used global measure for active learning systems, the area under the learning curve (ALC), was also reported in this study. The global ALC score was normalized by the area under the best achievable learning curve (1.00 ACC on all points of the learning curve). When measuring the ALC, two neighbor points on the curve were interpolated linearly.

To evaluate a pool-based active learning framework, we need not only a pool of unlabeled samples (that will be labeled during the querying step), but also an independently labeled test set. To generate reliable results, 10-fold CV was performed on active learning. At each CV iteration, nine folds formed the pool of unlabeled samples and the remaining fold was used for the evaluation of performance. For each ambiguous word in the MSH WSD dataset and a given querying algorithm (LC, Margin, or Entropy), 10 learning curves were generated from 10-fold CV experiments. Each learning curve started from two initial training samples and stopped at 90 training samples. An ALC was then created by averaging the ACC scores at each corresponding point for these 10 individual learning curves. The global score for each querying algorithm was then the ALC score from the averaged learning curve. Since the PL generated results with high variance due to random sampling, we averaged the results of the random querying method over 10 runs using the same start point, end point, and batch size.

To better summarize and compare the three querying methods and random sampling method, we generated a global ALC for each method from all learning curves of the 197 words in the WSD dataset. The global learning curve for a given method was generated by averaging points with the same number of training samples from all 197 ALCs.

To assess whether there is a significant difference between any two learners (three ALs and one PL) in terms of average ALC scores from 197 ambiguous words, we used the Wilcoxon signed rank test,[28] a non-parametric test for paired samples. As there were four different methods (six pair-wise comparisons in total), we applied a Bonferroni correction[29] to adjust for multiple comparisons, with family-wise type I error control at α=0.05. Therefore, if the p value from the Wilcoxon signed rank test was less than 0.0083 (0.05/6), we claimed that there was a statistically significant difference between two methods.

RESULTS

For each of 197 ambiguous words, we evaluated four learning methods (three ALs and one PL) and generated corresponding learning curves and global ALC scores. Table 1 shows the average ALC scores for all 197 words and some subsets. Detailed ALC scores for each ambiguous word and each learning algorithm are available in the online appendix (table A1). For any subsets in table 1, the three active learning algorithms had close average ALC scores, but they were better than the passive learning method (random sampling). Wilcoxon signed rank tests showed that the average ALC scores generated by ALs using the LC, Margin, or Entropy querying algorithms were statistically significantly better than ALC scores generated by the PL, in all subsets. However, the tests also revealed that the three ALs were not statistically significantly different. As shown in the last column of table 1, ALs outperformed the PL for 177 out of all 197 words (89.84%), 101 out of 111 abbreviations (90.99%), 76 out of 86 non-abbreviated terms (88.37%), and 13 out of 14 terms with more than two senses (92.85%).

Figure 1 shows the global learning curves across 197 words for the three active learning algorithms (LC, Margin, and Entropy) and the passive learning algorithm (Random). The learning curves of the three active learning algorithms almost overlapped, but they were clearly above the random sampling curve.

Based on the learning curves, we further reported the approximate numbers of training samples needed on average at different performance levels of supervised WSD systems for both active learning algorithms and random sampling (table 2).

Table 1 Average ALC scores for three active learning algorithms (Least Confidence (LC), Margin, and Entropy) and one passive learning method (Random), across all 197 ambiguous words and their subsets from the MSH WSD dataset

	Average ALC score				
MSH WSD dataset (subset)	LC	Margin	Entropy	Random	Active learner advantage percentage
197 Words	0.838	0.838	0.838	0.804	177 Out of 197 (89.84%)
111 Abbreviations	0.885	0.885	0.885	0.845	101 Out of 111 (90.99%)
86 Non-abbreviated terms	0.778	0.777	0.778	0.752	76 Out of 86 (88.37%)
14 Words with more than 2 senses	0.764	0.761	0.761	0.723	13 Out of 14 (92.85%)

ALC, area under the learning curve; LC, Least Confidence; WSD, word sense disambiguation.

We calculated the numbers of required training samples for different methods at different ACC values (0.75–0.90). It was clear that the ALs required fewer annotated training samples than the PL in order to reach the same accuracy for WSD tasks. For example, to train the WSD system to achieve an accuracy of 0.90, we needed 38 training samples for the random sampling method. But the ALs needed 24 training samples only, indicating a 37% (14/38) decrease in annotated training samples.

Furthermore, we also reported the performance of WSD systems integrated with different ALs and the PL when the number of training samples was fixed. Table 3 shows the accuracy of ALs and the PL when the number of training samples was set from 10 to 90, in increments of 10 samples. Our results showed that ALs could always generate higher accuracy than the PL when the same number of training samples was used. Additionally, the improvement of ALs was greater in the early stage (lower numbers of training samples needed).

As ALs may perform differently on multi-class classification tasks, we further conducted stratified analysis on the subset of 14 ambiguous words that had more than two senses. Table 4 (and online appendix tables A2 and A3) shows the detailed ALC scores of four learners for these individual words. ALs consistently showed better performance than the PL. Although LC achieved a slightly higher average ALC score (0.764) than Margin and Entropy (0.761), these differences were still not statistically significant according to the test. We also conducted stratified analysis on 111 abbreviations and detailed results can be found in the online appendix (tables A4 and A5). We noticed that abbreviations were relatively easier to disambiguate; ALs needed 50% fewer training samples on average than the PL (33 vs 67) in order to achieve an accuracy of 96%. This could be because acronyms are often accompanied by the expanded (unambiguous) forms, for example, 'extraction of acylcarnitine (AC) and amino acids (AA).' In addition, senses of the same abbreviation are generally quite unrelated, which probably makes the disambiguation task easier.

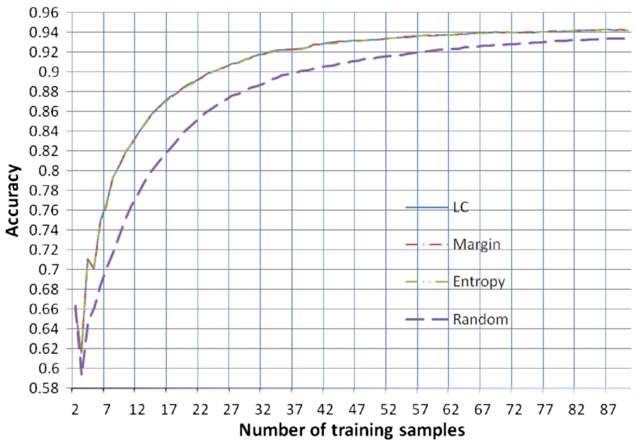

Figure 1 Global learning curves of 197 words in the MSH word sense disambiguation dataset for three active learning querying algorithms (LC, Margin, and Entropy) and the random querying method.

In addition, we tested the SVM-based WSD system alone by using all samples in the dataset, similar to the experiment in the previous study.[24] Our SVM-based WSD system achieved an average accuracy of 0.944 (via 10-fold CV) for all 197 words, which was similar to the previously reported result.[24]

DISCUSSION

In this study, we applied three different active learning algorithms to WSD tasks in the MEDLINE corpus. To the best of our knowledge, this is the first attempt to explore the use of active learning in supervised WSD tasks in the biomedical domain. Our results based on the MSH WSD dataset showed that WSD systems integrated with ALs significantly outperformed that with the PL (random sampling) in terms of average ALC score. Further analysis demonstrated that active learning strategies could not only reduce the number of training samples required for supervised WSD systems, but could also improve classification models when the same number of training samples was used. These findings suggest the great potential of active learning in improving the scalability of supervised WSD approaches in the biomedical domain. To achieve high performance on this dataset (over 90% accuracy), supervised WSD systems would require a few dozen sense-tagged instances for each ambiguous term when random sampling was used (table 2). By applying our current active learning strategies, we observed a reduction of 30–40% in annotation labor, which is promising. However, it is still not clear if such a reduction is good enough for building supervised WSD systems with a broad coverage, because the ambiguity problem is

Table 2 Approximate numbers of training samples needed on average at different accuracy values for both active learners and the passive learner

Accuracy	LC	Margin	Entropy	Random
0.70	5	5	5	7
0.75	6	6	6	10
0.80	9	9	9	15
0.85	13	13	13	21
0.90	24	24	24	38

LC, Least Confidence.

Table 3 Accuracy of active learners and the passive learner across 197 ambiguous words when different numbers of training samples were used

Number of training samples	LC	Margin	Entropy	Random
10	0.819	0.819	0.820	0.751
20	0.887	0.887	0.886	0.844
30	0.915	0.914	0.915	0.884
40	0.927	0.928	0.927	0.903
50	0.932	0.932	0.932	0.914
60	0.937	0.937	0.937	0.922
70	0.939	0.940	0.940	0.927
80	0.941	0.942	0.941	0.931
90	0.942	0.942	0.942	0.934

LC, Least Confidence.

pervasive in the biomedical domain. For example, Fundel and Zimmer[30] found that approximately 65% of 2.2 million human or rat related MEDLINE abstracts contained protein names that are ambiguous between the human and rat synonym lists. Liu *et al*[31] also reported that 33.1% of clinical abbreviations found in the UMLS 2001 distribution were ambiguous. In addition, annotation cost is also highly associated with the required performance of a task. If a WSD accuracy of 85% is good enough for a specific task, our active learning strategies would require only about a dozen sense-tagged instances on this dataset (see table 2). Therefore, a formal study is needed to further assess the feasibility of developing real-world WSD systems based on active learning, which should evaluate annotation costs at different levels of required performance.

We implemented three different querying algorithms for multi-class WSD tasks: LC, Margin, and Entropy. Although they are all uncertainty sampling-based algorithms, they are different when computing the uncertainty based on probabilities generated by the classifier: the LC algorithm considers the sense with the most confidence only; the Margin algorithm considers the two most likely senses; and the Entropy algorithm considers information for all possible senses. For the 14 words that had more than two senses in the dataset, we noticed a slight difference between LC and Margin/Entropy, but it was not statistically significant based on the statistical test, likely due to the small sample size (N=14). Another limitation of this study was that the pools for active learning were relatively small (maximum pool size was 500), as we used annotated samples in the MSH WSD dataset only. In a real-world application of active learning, we could collect a large number of unlabeled samples from MEDLINE for each ambiguous term, thus forming a much bigger pool for active learning experiments. We expect that larger pools will make the performance of active learning even better. We also noticed that most words in the MSH WSD dataset had almost equally distributed senses and only 17 out of 197 words had highly skewed senses. During the creation of the MSH WSD dataset, some minor senses were removed according to the procedure. In practice, imbalanced sense distribution will be observed more often, which could make WSD tasks more challenging.

Table 4 Active learning result for 14 words with more than two senses in the MSH word sense disambiguation (WSD) test collection

	Sense distribution					10-Fold CV ALC scores			
Words	S1	S2	S3	S4	S5	LC	Margin	Entropy	Random
Ala	98	97	98	0	0	0.825	0.812	0.824	0.762
Ca	89	98	98	98	0	0.615	0.620	0.601	0.580
Cold	93	96	62	0	0	0.686	0.683	0.683	0.668
Cortical	95	99	98	0	0	0.748	0.727	0.733	0.675
CP	97	99	98	0	0	0.868	0.864	0.866	0.822
DDS	99	98	20	0	0	0.827	0.829	0.828	0.772
Ice	98	37	98	0	0	0.755	0.759	0.762	0.759
Lens	97	99	99	0	0	0.716	0.681	0.689	0.662
Lupus	99	99	91	0	0	0.671	0.671	0.671	0.659
PCA	99	99	99	95	98	0.796	0.827	0.808	0.769
PCP	99	99	54	0	0	0.865	0.862	0.869	0.814
RA	99	99	99	0	0	0.869	0.872	0.872	0.795
TAT	99	99	99	0	0	0.719	0.714	0.711	0.672
THYMUS	99	96	99	0	0	0.735	0.734	0.743	0.711
Average						0.764	0.761	0.761	0.723

ALC, area under the learning curve; CV, cross validation; LC, Least Confidence.

We analyzed the learning curves of the 20 words where ALs did not perform better than the PL. We categorized the patterns of these cases as follows. (1) Poor models in the early stage: there was a cutoff point where the learning curves of AL and PL crossed over in the early stage of learning. AL performed poorly in the early stage before the cutoff but could outperform PL in the later stage. This pattern happened in 11 out of 20 cases. The reason could be that uncertainty sampling algorithms are sensitive to the quality of models. When the model is poor, the learning curve could be very unstable. The 'hasty generalization' problem pointed out by Wallace *et al*[32] could be one of the reasons for poor models in the early stage. Samples selected based on early uncertainty models may not be representative enough, especially for cases with skewed class distribution. As suggested by Wallace *et al*, one solution could be to apply diversity-based algorithms in the early stage. When the learning process passes the cutoff, active learning performs better than random because the classification model gets better. (2) Easy WSD cases: for some ambiguous words, high performance WSD models could be built based on only a small number of labeled samples. Basically they are easy WSD cases. For these cases, the informativeness or informative value of each sample is equally high and active learning is not necessary, as random sampling does the same job. We found three easy cases (*lymphogranulomatosis*, *PCD*, and *SLS*) out of 20 words. (3) Difficult WSD cases: this pattern was almost opposite to the second one. Even though we used all available samples with labels in the training set, the performance was not improved much. This indicates that the difference in informativeness or informative value among samples is small, and the informative value of each sample is equally low. We found three of these difficult cases (*Coffee*, *TMJ*, and *veterinary*). For the remaining three cases, the learning curves between AL and PL looked very similar. This could be due to the equal informativeness or informative value of each sample, or the querying algorithms failed to distinguish the difference in informativeness among unlabeled samples. These are also difficult cases because it is difficult to distinguish their samples.

Based on the above analysis, in order to further improve active learning for WSD tasks, we should investigate more robust active learning algorithms that can tolerate low quality models, or methods that can select good initial samples to build high quality models in the early stage. In addition to uncertainty sampling algorithms, other methods that consider different types of information (eg, sample diversity[33]) also need to be studied. We also plan to investigate other available WSD datasets that have more multiple senses so that we can test active learning algorithms on multi-class classification problems. Moreover, we are interested in applying active learning to real-world WSD tasks by developing an annotation interface that implements active learning querying algorithms for sample selection.

CONCLUSION

In this study, we integrated active learning algorithms with supervised WSD approaches to disambiguate terms in the MEDLINE corpus. Our evaluation using the MSH WSD dataset demonstrated that active learning strategies could not only reduce annotation cost but also improve the performance of supervised WSD models. In addition, we analyzed WSD cases where active learning algorithms did not achieve superior performance and provided useful insight for future improvements.

Funding This study was supported in part by grants from the US NIH: National Library of Medicine R01LM010681 and National Cancer Institute R01CA141307.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: the state of the art. *Comput Linguist* 1998;24:1–40.

2 Chen LF, Liu HF, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* 2005;21:248–56.

3 Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol* 2005;12:554–65.

4 Lee YK, Ng HT. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing—Volume 10*; Morristown, NJ, USA: Association for Computational Linguistics, 2002:41–8.

5 Conference on Empirical Methods in Natural Language Processing (EMNLP-96), Philadelphia, PA, May 1996:82–91.

6 Ng HT, Lee HB. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. *Proceedings of the 34th annual meeting on Association for Computational Linguistics*; Morristown, NJ, USA: Association for Computational Linguistics, 1996:40–7.

7 SENSEVAL '01 The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France, pp 123–6.

8 Bruce R, Wiebe J. Word-sense disambiguation using decomposable models. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*; Morristown, NJ, USA: Association for Computational Linguistics, 1994:139–46.

9 Liu HF, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Associ* 2004;11:320–31.

10 Yu H, Kim W, Hatzivassiloglou V, *et al*. Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *J Biomed Inform* 2007;40:150–9.

11 Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc* 2002;9:621–36.

12 Settles B. *Active learning literature survey*. University of Wisconsin-Madison, 2009, Computer Sciences Technical Report 1648.

13 Lewis DD, Gale WA. A sequential algorithm for training text classifiers. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*; 1994:3–12.

14 Tong S, Koller D. Support vector machine active learning with applications to text classification. *J Mach Learn Res* 2002;2:45–66.

15 Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2008:1069–78.

16 Tong S, Chang E. Support vector machine active learning for image retrieval. *Proceedings of the ACM International Conference on Multimedia*; 2001:107–18.

17 Chen J, Schein A, Ungar L, *et al*. An empirical study of the behavior of active learning for word sense disambiguation. *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*; 2006:120–7.

18 Kim S, Song Y, Kim K, *et al*. MMR-based active machine learning for bio named entity recognition. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*; 2006:69–72.

19 Tang M, Luo X, Roukos S. Active learning for statistical natural language parsing. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*; Philadelphia, July 2002:120–7.

20 Gangadharaiah R, Brown RD, Carbonell J. Active learning in example-based machine translation. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA*; 2009:227–30.

21 Zhu J, Hovy E. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*; 2007:783–90.

22 Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. *J Biomed Inform* 2012;45:265–72.

23 Figueroa RL, Zeng-Treitler Q, Ngo LH, *et al*. Active learning for clinical text classification: is it better than random sampling? *J Am Med Inform Assoc* 2012.

24 Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics* 2011;12:223.

25 NLM WSD Test Collection. http://wsd.nlm.nih.gov (accessed 21 Jan 2013).

26 Fan RE, Chang KW, Hsieh CJ, *et al*. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008;9:1871–4.

27 Guyon I, Cawley G, Dror G, *et al*. Results of the Active Learning Challenge. JMLR: Workshop and Conference Proceedings 2011;16:19–45.

28 Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945;1:80–3.

29 Hochberg Y, Tamhane AC. *Multiple comparison procedures*. New York: John Wiley & Sons, 1987.

30 Fundel K, Zimmer R. Gene and protein nomenclature in public databases. *BMC Bioinform* 2006;7:372.

31 Liu H, Lussier YA, Friedman C. A study of abbreviations in the UMLS. Proc AMIA Symp; 2001:393–7.

32 Wallace BC, Trikalinos TA, Lau J, *et al*. Semi-automated screening of biomedical citations for systematic reviews. *BMD Bioinform* 2010;11:55.

33 Kim S, Song Y, Kim K, *et al*. MMR-based active machine learning for bio named entity recognition. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*; 2006: 69–72.