An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation

[1] Jinying Chen, [1] Andrew Schein, [1] Lyle Ungar, [2] Martha Palmer

[1] Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104
{jinying,ais,ungar}@cis.upenn.edu

[2] Linguistic Department
University of Colorado
Boulder, CO, 80309
Martha.Palmer@colorado.edu

Abstract

This paper shows that two uncertainty-based active learning methods, combined with a maximum entropy model, work well on learning English verb senses. Data analysis on the learning process, based on both instance and feature levels, suggests that a careful treatment of feature extraction is important for the active learning to be useful for WSD. The overfitting phenomena that occurred during the active learning process are identified as classic overfitting in machine learning based on the data analysis.

1 Introduction

Corpus-based methods for word sense disambiguation (WSD) have gained popularity in recent years. As evidenced by the SENSEVAL exercises (http://www.senseval.org), machine learning models supervised by sense-tagged training corpora tend to perform better on the lexical sample tasks than unsupervised methods. However, WSD tasks typically have very limited amounts of training data due to the fact that creating large-scale high-quality sense-tagged corpora is difficult and time-consuming. Therefore, the lack of sufficient labeled training data has become a major hurdle to improving the performance of supervised WSD.

A promising method for solving this problem could be the use of active learning. Researchers use active learning methods to minimize the labeling of examples by human annotators. A decrease in overall labeling occurs because active learners (the machine learning models used in active learning) pick more informative examples for the target word (a word whose senses need to be learned) than those that would be picked randomly. Active learning requires human labeling of the newly selected training data to ensure high quality.

We focus here on *pool-based* active learning where there is an abundant supply of unlabeled data, but where the labeling process is expensive. In NLP problems such as text classification (Lewis and Gale, 1994; McCallum and Nigam, 1998), statistical parsing (Tang *et al.*, 2002), information extraction (Thompson *et al.*, 1999), and named entity recognition (Shen *et al.*, 2004), pool-based active learning has produced promising results.

This paper presents our experiments in applying two active learning methods, a min-margin based method and a Shannon-entropy based one, to the task of the disambiguation of English verb senses. The contribution of our work is not only in demonstrating that these methods work well for the active learning of coarse-grained verb senses, but also analyzing the behavior of the active learning process on two levels: the instance level and the feature level. The analysis suggests that a careful treatment of feature design and feature generation is important for a successful application of active learning to WSD. We also accounted for the overfitting phenomena that occurred in the learning process based on our data analysis.

The rest of the paper is organized as follows. In Section 2, we introduce two uncertainty sampling methods used in our active learning experiments and review related work in using active learning for WSD. We then present our active learning experiments on coarse-grained English verb senses in Section 3 and analyze the active learning

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 120–127,
New York, June 2006. ©2006 Association for Computational Linguistics

process in Section 4. Section 5 presents conclusions of our study.

2 Active Learning Algorithms

The methods evaluated in this work fit into a common framework described by Algorithm 1 (see Table 1). The key difference between alternative active learning methods is how they assess the value of labeling individual examples, i.e., the methods they use for ranking and selecting the candidate examples for labeling. The framework is wide open to the type of ranking rule employed. Usually, the ranking rule incorporates the model trained on the currently labeled data. This is the reason for the requirement of a partial training set when the algorithm begins.

```
Algorithm 1
Require: initial training set, pool of unlabeled examples
Repeat
    Select $T$ random examples from pool
    Rank $T$ examples according to active learning rule
    Present the top-ranked example to oracle for labeling
    Augment the training set with the new example
Until Training set reaches desirable size
```

Table 1. A Generalized Active Learning Loop

In our experiments we look at two variants of the *uncertainty sampling* heuristic: entropy sampling and margin sampling. Uncertainty sampling is a term invented by Lewis and Gale (Lewis and Gale, 1994) to describe a heuristic where a probabilistic classifier picks examples for which the model's current predictions are least certain. The intuitive justification for this approach is that regions where the model is uncertain indicate a decision boundary, and clarifying the position of decision boundaries is the goal of learning classifiers. Schein (2005) demonstrates the two methods run quickly and compete favorably against alternatives when combined with the logistic regression classifier.

2.1 Entropy Sampling

A key question is how to measure uncertainty. Different methods of measuring uncertainty will lead to different variants of uncertainty sampling. We will look at two such measures. As a convenient notation we use **q** (a vector) to represent the trained model's predictions, with q_c equal to the predicted probability of class c. One method is to pick the example whose prediction vector **q** displays the greatest Shannon entropy:

$$-\sum_c q_c \log q_c \qquad (1)$$

Such a rule means ranking candidate examples in Algorithm 1 by Equation 1.

2.2 Margin Sampling

An alternative method picks the example with the smallest margin: the difference between the largest two values in the vector **q** (Abe and Mamitsuka, 1998). In other words, if c and c' are the two most likely categories for example x_n, the margin is measured as follows:

$$M_n = \left|\Pr(c \mid x_n) - \Pr(c' \mid x_n)\right| \qquad (2)$$

In this case Algorithm 1 would rank examples by increasing values of margin, with the smallest value at the top of the ranking.

Using either method of uncertainty sampling, the computational cost of picking an example from T candidates is: *O(TD)* where D is the number of model parameters.

2.3 Related Work

To our best knowledge, there have been very few attempts to apply active learning to WSD in the literature (Fujii and Inui, 1999; Chklovski and Mihalcea, 2002; Dang, 2004). Fujii and Inui (1999) developed an example sampling method for their example-based WSD system in the active learning of verb senses in a pool-based setting. Unlike the uncertainty sampling methods (such as the two methods we used), their method did not select examples for which the system had the minimal certainty. Rather, it selected the examples such that after training using those examples the system would be most certain about its predictions on the rest of the unlabeled examples in the next iteration. This sample selection criterion was enforced by calculating a training utility function. The method performed well on the active learning of Japanese verb senses. However, the efficient computation of the training utility function relied on the nature of the example-based learning method, which made their example sampling method difficult to export to other types of machine learning models.

Open Mind Word Expert (Chklovski and Mihalcea, 2002) was a real application of active learning for WSD. It collected sense-annotated examples from the general public through the Web to create the training data for the SENSEVAL-3 lexical sample tasks. The system used the disagreement of two classifiers (which employed different sets of features) on sense labels to evaluate the difficulty of the unlabeled examples and ask the web users to tag the difficult examples it selected. There was no formal evaluation for this active learning system.

Dang (2004) used an uncertainty sampling method to get additional training data for her WSD system. At each iteration the system selected a small set of examples for which it had the lowest confidence and asked the human annotators to tag these examples. The experimental results on 5 English verbs with fine-grained senses (from WordNet 1.7) were a little surprising in that active learning performed no better than random sampling. The proposed explanation was that the quality of the manually sense-tagged data was limited by an inconsistent or unclear sense inventory for the fine-grained senses.

3 Active Learning Experiments

3.1 Experimental Setting

We experimented with the two uncertainty sampling methods on 5 English verbs that had coarse-grained senses (see Table 2), as described below. By using coarse-grained senses, we limit the impact of noisy data due to unclear sense boundaries and therefore can get a clearer observation of the effects of the active learning methods themselves.

verb	# of sen.	baseline acc. (%)	Size of data for active learning	Size of test data
Add	3	91.4	400	100
Do	7	76.9	500	200
Feel	3	83.6	400	90
See	7	59.7	500	200
Work	9	68.3	400	150

Table 2. The number of senses, the baseline accuracy, the number of instances used for active learning and for held-out evaluation for each verb

The coarse-grained senses are produced by grouping together the original WordNet senses using syntactic and semantic criteria (Palmer *et al.*, 2006). Double-blind tagging is applied to 50 instances of the target word. If the ITA < 90%, the sense entry is revised by adding examples and explanations of distinguishing criteria.

Table 2 summarizes the statistics of the data. The baseline accuracy was computed by using the "most frequent sense" heuristic to assign sense labels to verb instances (examples). The data used in active learning (Column 4 in Table 2) include two parts: an initial labeled training set and a pool of unlabeled training data. We experimented with sizes 20, 50 and 100 for the initial training set. The pool of unlabeled data had actually been annotated in advance, as in most pool-based active learning experiments. Each time an example was selected from the pool by the active learner, its label was returned to the learner. This simulates the process of asking human annotators to tag the selected unlabeled example at each time. The advantage of using such a simulation is that we can experiment with different settings (different sizes of the initial training set and different sampling methods).

The data sets used for active learning and for held-out evaluation were randomly sampled from a large data pool for each round of the active learning experiment. We ran ten rounds of the experiments for each verb and averaged the learning curves for the ten rounds.

In the experiments, we used random sampling (picking up an unlabeled example randomly at each time) as a lower bound. Another control (ultimate-maxent) was the learner's performance on the test set when it was trained on a set of labeled data that were randomly sampled from a large data pool and equaled the amount of data used in the whole active learning process (e.g., 400 training data for the verb *add*).

The machine learning model we used for active learning was a regularized maximum entropy (MaxEnt) model (McCallum, 2002). The features used for disambiguating the verb senses included topical, collocation, syntactic (e.g., the subject, object, and preposition phrases taken by a target verb), and semantic (e.g., the WordNet synsets and hypernyms of the head nouns of a verb's NP arguments) features (Chen and Palmer, 2005).

3.2 Experimental Results

Due to space limits, Figure 1 only shows the learning curves for 4 verbs *do, feel, see,* and *work* (size of the initial training set = 20). The curve for the verb *add* is similar to that for *feel*. These curves clearly show that the two uncertainty sampling methods, the entropy-based (called entropy-maxent in the figure) and the margin-based (called min_margin-maxent), work very well for active learning of the senses of these verbs.

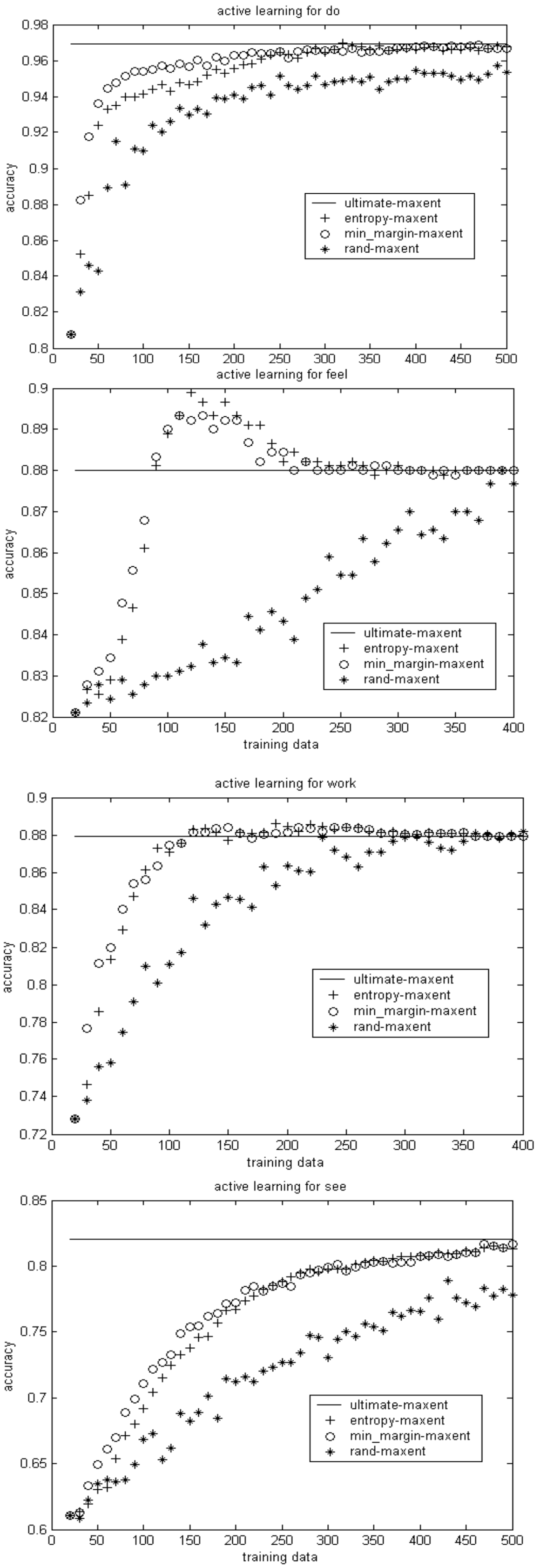

Figure 1 Active learning for four verbs

Both methods outperformed the random sampling method in that they reached the upper-bound accuracy earlier and had smoother learning curves. For the four verbs *add, do, feel* and *see*, their learning curves reached the upper bound at about 200~300 iterations, which means 1/2 or 1/3 of the annotation effort can be saved for these verbs by using active learning, while still achieving the same level of performance as supervised WSD without using active learning. Given the large-scale annotation effort currently underway in the OntoNotes project (Hovy *et al.*, 2006), this could provide considerable savings in annotation effort and speed up the process of providing sufficient data for a large vocabulary. The OntoNotes project has now provided coarse-grained entries for over 350 verbs, with corresponding double–blind annotation and adjudication in progress. As this adjudicated data becomes available, we will be able to train our system accordingly. Preliminary results for 22 of these coarse-grained verbs (with an average grouping polysemy of 4.5) give us an average accuracy of 86.3%. This will also provide opportunities for more experiments with active learning, where there are enough instances. Active learning could also be beneficial in porting these supervised taggers to new genres with different sense distributions.

We also experimented with different sizes of the initial training set (20, 50 and 100) and found no significant differences in the performance at different settings. That means, for these 5 verbs, only 20 labeled training instances will be enough to initiate an efficient active learning process.

From Figure 1, we can see that the two uncertainty sampling methods generally perform equally well except that for the verb *do*, the min-margin method is slightly better than the entropy method at the beginning of active learning. This may not be so surprising, considering that the two methods are equal for two-class classification tasks (see Equations 1 and 2 for their definition) and the verbs used in our experiments have coarse-grained senses and often have only 2 or 3 major senses.

An interesting phenomenon observed from these learning curves is that for the two verbs *add* and *feel*, the active learner reached the upper bound very soon (at about 100 iterations) and then even breached the upper bound. However, when the training set was extended, the learner's performance dropped and eventually returned to the same level of the upper bound. We discuss the phenomenon below.

4 Analysis of the Learning Process

In addition to verifying the usefulness of active learning for WSD, we are also interested in a deeper analysis of the learning process. For example, why does the active learner's performance drop sometimes during the learning process? What are the characteristics of beneficial features that help to boost the learner's accuracy? How do we account for the overfitting phenomena that occurred during the active learning for the verbs *add* and *feel*? We analyzed the effect of both instances and features throughout the course of active learning using min-margin-based sampling.

4.1 Instance-level Analysis

Intuitively, if the learner's performance drops after a new example is added to the training set, it is likely that something has gone wrong with the new example. To find out such *bad* examples, we define a measure *credit_inst* for instance *i* as:

$$\frac{1}{m}\sum_{r=1}^{m}\sum_{l=1}^{n} sel(i,l)(Acc_{l+1} - Acc_{l}) \qquad (3)$$

where Acc_l and Acc_{l+1} are the classification accuracies of the active learner at the *lth* and *(l+1)th* iterations. *n* is the total number of iterations of active learning and *m* is the number of rounds of active learning (*m*=10 in our case). $sel(i,l)$ is 1 *iff* instance *i* is selected by the active learner at the *l*th iteration and is 0 if otherwise.

An example is a *bad example* if and only if it satisfies the following conditions:

a) its *credit_inst* value is negative

b) it increases the learner's performance, if it does, less often than it decreases the performance in the 10 rounds.

We ranked the bad examples by their *credit_inst* values and their frequency of decreasing the learner's performance in the 10 rounds. Table 3 shows the top five bad examples for *feel* and *work*. There are several reasons why the bad examples may hurt the learner's performance. Column 3 of Table 3 proposes reasons for many of our bad examples. We categorized these reasons into three major types.

I. The major senses of a target verb depend heavily on the semantic categories of its NP arguments but WordNet sometimes fails to provide the appropriate semantic categories (features) for the head nouns of these NP arguments. For example, *feel* in *the board apparently felt no pressure* has Sense 1 (experience). In Sense 1, *feel* typically takes an ***animate*** subject. However, *board*, the head word of the verb's subject in the above sentence has no animate meanings defined in WordNet. Even worse, the major meaning of *board*, i.e., *artifact*, is typical for the subject of *feel* in Sense 2 (touch, grope). Similar semantic type mismatches hold for the last four bad examples of the verb *work* in Table 3.

II. The contexts of the target verb are difficult for our feature exaction module to analyze. For example, the antecedent for the pronoun subject *they* in the first example of *work* in Table 3 should be *ringers*, an ***agent*** subject that is typical for Sense 1 (exert oneself in an activity). However, the feature exaction module found the wrong antecedent *changes* that is an unlikely fit for the intended verb sense. In the fourth example for *feel*, the feature extraction module cannot handle the expletive "it" (a dummy subject) in "it was felt that", therefore, it cannot identify the typical syntactic pattern for Sense 3 (find, conclude), i.e., *subject+feel+relative clause*.

III. Sometimes, deep semantic and discourse analyses are needed to get the correct meaning of the target verb. For example, in the third example of *feel*, "*…, he or she feels age creeping up*", it is difficult to tell whether the verb has Sense 1 (experience) or Sense 3 (find) without an understanding of the meaning of the relative clause and without looking at a broader discourse context. The syntactic pattern identified by our feature extraction module, *subject+feel+relative clause*, favors Sense 3 (find), which leads to an inaccurate interpretation for this case.

Recall that the motivation behind uncertainty samplers is to find examples near decision boundaries and use them to clarify the position of these boundaries. Active learning often does find informative examples, either ones from the less common senses or ones close to the boundary between the different senses. However, active learning also identifies example sentences that are difficult to analyze. The failure of our feature extraction module, the lack of appropriate semantic categories for certain NP arguments in WordNet, the lack of deep analysis (semantic and discourse analysis) of the context of the target verb can all

feel	**Proposed reasons for bad examples**	**Senses**
Some days the coaches make you feel as though you are part of a large herd of animals .	?	S1: experience
And , with no other offers on the table , the board apparently felt no pressure to act on it.	subject: *board*, no “animate” meaning in WordNet	S1: experience
Sometimes a burst of aggressiveness will sweep over a man -- or his wife -- because he or she feels age creeping up.	syntactic pattern: sbj+*feel*+relative clause headed by *that*, a typical pattern for Sense 3 (find) rather than Sense 1 (experience)	S1: experience
At this stage it was felt I was perhaps more pertinent as chief. executive .	syntactic pattern: sbj+*feel*+relative clause, typical for Sense 3 (find) but has not been detected by the feature exaction module	S3: find, conclude
I felt better Tuesday evening when I woke up.	?	S1: experience
Work		
When their changes are completed, and after they have worked up a sweat, ringers often ……	subject: *they*, the feature exaction module found the wrong antecedent (*changes* rather than *ringers*) for *they*	S1: exert oneself in an activity
Others grab books, records , photo albums , sofas and chairs , working frantically in the fear that an aftershock will jolt the house again .	subject: *others* (means *people* here), no definition in WordNet	S1: exert oneself in an activity
Security Pacific 's factoring business works with companies in the apparel, textile and food industries …	subject: *business*, no “animate” meaning in WordNet	S1: exert oneself in an activity
… ; blacks could work there , but they had to leave at night .	subject: *blacks*, no “animate” meaning in WordNet	S1: exert oneself in an activity
… has been replaced by alginates (gelatin-like material) that work quickly and accurately and with least discomfort to a child .	subject: *alginates*, unknown by WordNet	S2: perform, function, behave

Table 3 Data analysis of the top-ranked bad examples found for two verbs

produce misleading features. Therefore, in order to make active learning useful for its applications, both identifying difficult examples *and* getting good features for these examples are equally important. In other words, a careful treatment of feature design and feature generation is necessary for a successful application of active learning.

There is a positive side to identifying such “bad” examples; one can have human annotators look at the features generated from the sentences (as we did above), and use this to improve the data or the classifier. Note that this is exactly what we did above: the identification of bad sentences was automatic, and they could then be reannotated or removed from the training set or the feature extraction module needs to be refined to generate informative features for these sentences.

Not all sentences have obvious interpretations; hence the two question marks in Table 3. An example can be bad for many reasons: conflicting features (indicative of different senses), misleading features (indicative of non-intended senses), or just containing random features that are incorrectly incorporated into the model. We will return to this point in our discussion of the overfitting phenomena for active learning in Section 4.3.

4.2 Feature-level Analysis

The purpose of our feature-level analysis is to identify informative features for verb senses. The learning curve of the active learner may provide some clues. The basic idea is, if the learner’s performance increases after adding a new example, it is likely that the *good* example contains good features that contribute to the clarification of sense boundaries. However, the feature-level analysis is much less straightforward than the instance-level analysis since we cannot simply say the features that are active (present) in this *good* example are all good. Rather, an example often contains both good and bad features, and many other features that are somehow neutral or uninformative. The interaction or balance between these features determines the final outcome. On the other hand, a statistics based analysis may help us to find features that tend to be good or bad. For this analysis, we define a measure *credit_feat* for feature *i* as:

$$\frac{1}{m}\sum_{r=1}^{m}\sum_{l=1}^{n} active(i,l)(Acc_{l+1} - Acc_{l})\frac{1}{act_{l}} \qquad (4)$$

where $active(i,l)$ is 1 iff feature i is active in the example selected by the active learner at the lth iteration and is 0 if otherwise. act_l is the total number of active features in the example selected at the lth iteration. n and m have the same definition as in Equation 3.

A feature is regarded as *good* if its *credit_feat* value is positive. We ranked the good features by their *credit_feat* values. By looking at the top-ranked good features for the verb *work* (due to space limitations, we omit the table data), we identify two types of typically good features.

The first type of good feature occurs frequently in the data and has a frequency distribution over the senses similar to the data distribution over the senses. Such features include those denoting that the target verb takes a subject (*subj*), is not used in a passive mode (*morph_normal*), does not take a direct object (*intransitive*), occurs in present tense (*word_work, pos_vb, word_works, pos_vbz*), and semantic features denoting an ***abstract*** subject (subjsyn_16993 [1]) or an ***entity*** subject (subjsyn_ 1742), *etc.* We call such features *background* features. They help the machine learning model learn the appropriate sense distribution of the data. In other words, a learning model only using such features will be equal to the “most frequent sense” heuristic used in WSD.

Another type of good feature occurs less frequently and has a frequency distribution over senses that mismatches with the sense distribution of the data. Such features include those denoting that the target verb takes an inanimate subject (*subj_it*), takes a particle *out* (*prt_out*), is followed directly by the word *out* (word+1_out), or occurs at the end of the sentence. Such features are indicative of less frequent verb senses that still occur fairly frequently in the data. For example, taking an ***inanimate*** subject (*subj_it*) is a strong clue for Sense 2 (perform, function, behave) of the verb *work*. Occurring at the end of the sentence is also indicative of Sense 2 since when *work* is used in Sense 1 (exert oneself in an activity), it tends to take adjuncts to modify the activity as in *He is working hard to bring up his grade*.

There are some features that don’t fall into the above two categories, such as the topical feature *tp_know* and the collocation feature *pos-2_nn*. There are no obvious reasons why they are good for the learning process, although it is possible that the combination of two or more such features could make a clear sense distinction. However, this hypothesis cannot be verified by our current statistics-based analysis. It is also worth noting that our current feature analysis is post-experimental (i.e., based on the results). In the future, we will try automatic feature selection methods that can be used in the training phase to select useful features and/or their combinations.

We have similar results for the feature analysis of the other four verbs.

4.3 Account for the Overfitting Phenomena

Recall that in the instance-level analysis in Section 4.1, we found that some examples hurt the learning performance during active learning but for no obvious reasons (the two examples marked by ? in Table 3). We found that these two examples occurred in the overfitting region for *feel*. By looking at the bad examples (using the same definition for *bad* example as in Section 4.1) that occurred in the overfitting region for both *feel* and *add*, we identified two major properties of these examples. First, most of them occurred only once as bad examples (19 out 23 for *add* and 40 out of 63 for *feel*). Second, many of the examples had no obvious reasons for their badness.

Based on the above observations, we believe that the overfitting phenomena that occurred for the two verbs during active learning is typical of classic overfitting, which is consistent with a *"death by a thousand mosquito bites"* of rare bad features, and consistent with there often being (to mix a metaphor) no *"smoking gun"* of a bad feature/instance that is added in, especially in the region far away from the starting point of active learning.

5 Conclusions

We have shown that active learning can lead to substantial reductions (often by half) in the number of observations that need to be labeled to achieve a given accuracy in word sense disambiguation, compared to labeling randomly selected instances. In a follow-up experiment, we also compared a larger number of different active learning methods.

[1] Those features are from the WordNet. The numbers are WordNet ids of synsets and hypernyms.

The results suggest that for tasks like word sense disambiguation where maximum entropy methods are used as the base learning models, the minimum margin active criterion for active learning gives superior results to more comprehensive competitors including bagging and two variants of query by committee (Schein, 2005). By also taking into account the high running efficiency of the min-margin method, it is a very promising active learning method for WSD.

We did an analysis on the learning process on two levels: instance-level and feature-level. The analysis suggests that a careful treatment of feature design and feature generation is very important for the active learner to take advantage of the difficult examples it finds during the learning process. The feature-level analysis identifies some characteristics of good features. It is worth noting that the good features identified are not particularly tied to active learning, and could also be obtained by a more standard feature selection method rather than by looking at how the features provide benefits as they are added in.

For a couple of the verbs examined, we found that active learning gives higher prediction accuracy midway through the training than one gets after training on the entire corpus. Analysis suggests that this is not due to *bad* examples being added to the training set. It appears that the widely used maximum entropy model with Gaussian priors is overfitting: the model by including too many features and thus fitting noise as well as signal. Using different strengths of the Gaussian prior does not solve the problem. If a very strong prior is used, then poorer accuracy is obtained. We believe that using appropriate feature selection would cause the phenomenon to vanish.

Acknowledgements

This work was supported by National Science Foundation Grant NSF-0415923, Word Sense Disambiguation, the DTO-AQUAINT NBCHC-040036 grant under the University of Illinois subcontract to University of Pennsylvania 2003-07911-01 and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the DTO, or DARPA.

References

Naoki Abe and Hiroshi Mamitsuka. 1998. Query learning strategies using boosting and bagging. In *Proc. of ICML1998*, pages 1–10.

Jinying Chen and Martha Palmer. 2005. Towards Robust High Performance Word Sense Disambiguation of English Verbs Using Rich Linguistic Features, In *Proc. of IJCNLP2005*, Oct., Jeju, Republic of Korea.

Tim Chklovski and Rada Mihalcea, Building a Sense Tagged Corpus with Open Mind Word Expert, in *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Philadelphia, July 2002.

Hoa T. Dang. 2004. Investigations into the role of lexical semantics in word sense disambiguation. PhD Thesis. University of Pennsylvania.

Atsushi Fujii, Takenobu Tokunaga, Kentaro Inui, Hozumi Tanaka. 1998. Selective sampling for example-based word sense disambiguation, Computational Linguistics, v.24 n.4, p.573-597, Dec.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. OntoNotes: The 90% Solution. Accepted by *HLT-NAACL06*. Short paper.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In W. Bruce Croft and Cornelis J. van Rijsbergen, editors, *Proceedings of SIGIR-94*, Dublin, IE.

Andrew K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. *http://www.cs.umass.edu/~mccallum/mallet*.

Andew McCallum and Kamal Nigam. 1998. Employing EM in pool-based active learning for text classification. In *Proc. of ICML '98*.

Martha Palmer, Hoa Trang Dang and Christiane Fellbaum. (to appear, 2006). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering.*

Andrew I. Schein. 2005. Active Learning for Logistic Regression. Ph.D. Thesis. Univ. of Pennsylvania.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou and Chew Lim Tan. 2004 Multi-criteria-based active learning for named entity recognition, In *Proc. of ACL04*, Barcelona, Spain.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proc. of ACL* 2002.

Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proc. of ICML-99*.