Using Variance as a Stopping Criterion for Active Learning of Frame Assignment

Masood Ghayoomi
German Grammar Group
Freie Universität Berlin
Berlin, 14195
masood.ghayoomi@fu-berlin.de

Abstract

Active learning is a promising method to reduce human's effort for data annotation in different NLP applications. Since it is an iterative task, it should be stopped at some point which is optimum or near-optimum. In this paper we propose a novel stopping criterion for active learning of frame assignment based on the variability of the classifier's confidence score on the unlabeled data. The important advantage of this criterion is that we rely only on the unlabeled data to stop the data annotation process; as a result there are no requirements for the gold standard data and testing the classifier's performance in each iteration. Our experiments show that the proposed method achieves 93.67% of the classifier maximum performance.

1 Introduction

Using supervised machine learning methods is very popular in Natural Language Processing (NLP). However, these methods are not applicable for most of the NLP tasks due to the lack of labeled data. Although a huge amount of unlabeled data is freely available, labeling them for supervised learning techniques is very tedious, expensive, time consuming, and error prone.

Active learning is a supervised machine learning method in which informative instances are chosen by the classifier for labeling. Unlike the normal supervised set-up where data annotation and learning are completely independent, active learning is a sequential process (Settles, 2009; Busser and Morante, 2005). This learning method is used in a variety of NLP tasks such as information extraction (Thompson et al., 1999), semantic role labeling (Busser and Morante, 2005), machine translation (Haffari and Sarkar, 2009), and name entity recognition (Laws and Schütze, 2008). In our study, we apply this method for the frame assignment task as a kind of semantic analysis.

The process of active learning is as follows: the learner takes a set of labeled instances, called *seed* data, as an input for initial training of the classifier; and then a larger set of unlabeled instances will be selected by the classifier to be labeled with the human interaction. Even a small set of well selected samples for labeling can achieve the same level of performance of a large labeled data set; and the oracle's effort will be reduced as a result.

The motivation behind active learning is selecting the most useful examples for the classifier and thereby minimizing the annotation effort while still keeping up the performance level (Thompson et al., 1999). There are two major learning scenarios in active learning which are very popular among researchers and frequently used in various NLP tasks: *stream-based sampling* (Cohn et al., 1994) and *pool-based sampling* (Lewis and Gale, 1994).

The samples that are selected should be hard and very informative. There are different query methods for sample selection which are independent of the active learning scenarios (Settles, 2009). Among them, *uncertainty sampling* (Lewis and Gale, 1994) is the most well-known and the simplest sample selection method which only needs one classifier (Baldridge and Osborne, 2004). In this query method, the samples that the classifier is least con-

Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing, pages 1–9,
Los Angeles, California, June 2010.

Algorithm 1 Uncertainty Sampling in Active Learning

Input: Seed data S, Pool of unlabeled samples U
Use S to train the classifier C
while the stopping criterion is met **do**
  Use C to annotate U
  Select the top K samples from U predicted by C which have the lowest confidence
  Label K, augment S with the$K$ samples, and remove K from U
  Use S to retrain C
end while

fident on their labels are selected and handed out to the oracle. To this aim, a confidence score is required which is in fact the prediction of the classifier with the highest probability for the label of the sample (Busser and Morante, 2005).

The approach taken in active learning for our task is based on the uncertainty of the classifier with access to the pool of data. The learning process is presented in Algorithm 1. Since active learning is an iterative process (Busser and Morante, 2005), it should be stopped at some point which is optimum or at least near-optimum. A learning curve is used as a means to illustrate the learning progress of the learner, so that we can monitor the performance of the classifier. In fact, the curve signals when the learning process should stop as almost no increase or even a drop in the performance of the classifier is observed. At this point, additional training data will not increase the performance any more. In this paper, we propose a new stopping criterion based on the variability of the classifier's confidence score on the selected unlabeled data so that we avoid using the labeled gold standard.

The structure of the paper is as follows. In Section 2, we briefly describe *frame semantics* as it is the domain of application for our model. Section 3 introduces our stopping criterion and describes the idea behind it. In Section 4, we describe our data set and present the experimental results. In Section 5, related work on stopping criteria is outlined; and finally Section 6 summarizes the paper.

2 Frame Semantics

Syntactic analysis such as part-of-speech (POS) tagging and parsing has been widely studied and has achieved a great progress. However, semantic analysis did not have such a rapid progress. This problem has recently motivated researches to pay special attention to natural language understanding since it is one of the essential parts in information extraction and question-answering.

Frame semantic structure analysis which is based on the case grammar of Fillmore (1968) is one of the understanding techniques to provide the knowledge about the actions, the participants of the action, and the relations between them. In Fillmore's view, a frame is considered as an abstract scene having some participants as the arguments of the predicate, and some sentences to describe the scene. In fact the frames are the conceptual structures for the background knowledge of the abstract scenes represented by the lexical units and provide context to the elements of the action. FrameNet (Baker and Lowe, 1998) is a data set developed at ICSI Berkley University based on the frame semantics.

In frame semantic structure analysis, the semantic roles of the elements participating in the action are identified. Determining and assigning the semantic roles automatically require two steps: *frame assignment*, and *role assignment* (Erk and Pado, 2006). The first step consists in identifying the frame which is evoked by the predicate to determine the unique frame that is appropriate for the sample. The next step is identifying the arguments of the predicate and assigning the semantic roles to the syntactic arguments of the given frame. In our research, we study the first step, and leave the second step for future work.

3 The Proposed Stopping Criterion

The main idea behind the stopping criteria is to stop the classifier when it has reached its maximum performance and labeling of further examples from the unlabeled data set will not increase the classifier's performance any more. Determining this point is very difficult experimentally without access to the gold standard labels to evaluate the performance; however, we should find a criterion to stop active learning in a near-optimum point. To this aim, we propose a novel stopping criterion which uses the *variance* of the classifier's confidence score for the predicted labels to represent the degree of spreading out the confidence scores around their mean. We hypothesize that there is a correlation between the

performance saturation of the classifier and the variability on the confidence of the selected instances.

Generally, as we will see in Section 5, a stopping criterion could be based either on the performance of the classifier on the test data, or on the confidence score of the classifier on the unlabeled data. In our method, we used the second approach. The biggest advantage of this model is that no gold standard data is required to evaluate the performance of the system in each iteration.

3.1 Mean and Variance

Mean and variance are two of the well-known statistical metrics. Mean is a statistical measurement for determining the central tendency among a set of scores. In our study, we have computed the mean (*M*) of the classifier's confidence score for the predicted labels of 5 samples selected in each iteration. Variance is the amount of variability of the scores around their mean. To compute the variability of the classifier's confidence score for the selected samples in each iteration, the following equation is used in our task:

$$Variance = \frac{\sum_{i=1}^{K}(C_i - M)^2}{K} \quad (1)$$

where C_i is the confidence score of each selected sample in each iteration, *M* is the mean of the confidence scores for the predicted labels, and *K* is the number of samples selected in the same iteration (*K*=5 in our study).

3.2 The General Idea

According to the pool-based scenario, in each iteration *K* samples of the extra unlabeled data which have the lowest confidence score are selected, and after labeling by the oracle they are added to the training data. In the early iterations, the mean of the classifier's confidence score for the selected samples is low. Since the classifier is not trained enough in these iterations, most of the scores are low and they do not have a high degree of variability. As a result the variance of the confidence score for these samples is low. We call this step the *untrained* stage of the classifier.

As the classifier is training with more data, the confidence score of the samples will gradually increase; as a result, there will be a high degree of variability in the confidence scores which spread out around their mean. In these iterations, the classifier is relatively in the borderline of the training stage, passing from untrained to trained; so that there will be a high variability of confidence scores which leads to have a high variance. This is the *training* stage of the classifier.

When the classifier is trained, the confidence score of the classifier on the selected samples will increase. However, from a certain point that the classifier is trained enough, all of the confidence scores are located tightly around their mean with a low degree of variability; as a result, the variance of the samples decreases. This is the stage that the classifier is *trained*.

The curve in Figure 1 represents the behavior of the variance in different iterations such that the *x* axis is the number of iterations, and the *y* axis is the variance of the confidence scores in each iteration.

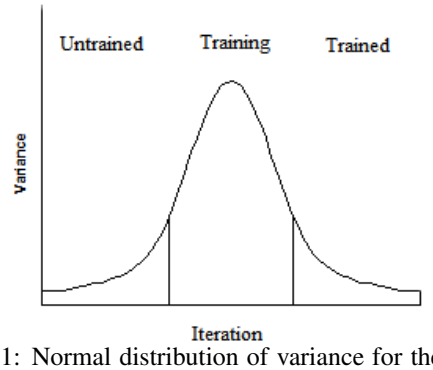

Figure 1: Normal distribution of variance for the classifier's confidence score

Based on our assumption, the best stopping point is when variance reaches its global peak and starts to decrease. In this case, the classifier passes the training stage and enters into the trained stage.

3.3 The Variance Model

It is difficult to determine the peak of the variance on the fly, i.e. without going through all iterations. One easy solution is to stop the learning process as soon as there is a decrease in the variance. However, as it is very likely to stick in the local maxima of the variance curve, this criterion does not work well. In other words, it is possible to have small peaks before reaching the global peak, the highest variability of the classifier's confidence score; so that we might stop at some point we are not interested in and it should be ignored.

To avoid this problem, we propose a model, called *variance model* (VM), to stop active learning when variance (V) decreases in *n* sequential iterations; i.e.

$$V_i < V_{i-1} < ... < V_{i-n}.$$

There is a possibility that this condition is not satisfied at all in different iterations. In such cases, active learning will not stop and all data will be labeled. This condition is usually met when there are instances in the data which are inherently ambiguous. Having such data is generally unavoidable and it is often problematic for the learner.

Although the above model can deal with the local maxima problem, there is a possibility that the decreased variance in *n* sequential iterations is very small and it is still possible to stick in the local maxima. To avoid this problem and have a better stopping point, we extend the proposed model by setting a threshold *m*, called the *Extended Variance Model* (EVM), in which the minimum variance decrement in *n* sequential iterations must be *m*; i.e.

$$V_i < V_{i-1} - m < ... < V_{i-n} - m.$$

4 Experimental Results

4.1 Setup of Experiment

What we aim to do in our study is assigning frames with active learning. We have chosen the pool-based scenario by using the uncertainty sampling method. In our task, since we have a small data set, 5 instances (*K*=5) with the lowest confidence score of the predited labels will be selected in each iteration from the pool of data and handed out to the oracle to be labeled.

We have used a toolkit for the supervised word sense disambiguation task called *Majo* (Rehbein et al., 2009) which has a graphical user interface (GUI) for semantic annotation based on active learning. The toolkit supports German and English; and it uses the openNLP MAXENT package[1] to build the model. In this toolkit, the confidence score of the classifier is the posterior probability of the most probable label assigned to each sample.

In addition, there are some built-in plugins in the tool for syntactic and semantic pre-processing to provide the relevant features for the classifier. We utilized the following plugins that support English:

- *Stanford Word Range Plugin* provides features based on the local context of the surface string for the target. The window size of the local context can be set manually in the GUI. Based on initial experiments for the target verbs, we found out that a window ± 3 performs the best.
- *Stanford POS Tag Word Range Plugin* provides the POS tags of the words within a sentence by using Stanford POS Tagger. In this plugin, the window size could also be set manually to extract the POS local context of the target word. Based on initial experiments, a window of ± 3 achieved the best performance.
- *Berkley Sentence Phrase Plugin* utilizes the Berkley Parser and provides the syntactic analysis of the sentence. This plugin is used to extract all word forms of the children nodes from a particular syntactic mother node (VP in our study) and add them to the feature set.
- *Berkley Sentence Phrase POS Tag Plugin* uses the Berkley POS tagger such that we define the mother node of the target word in the parse tree (VP in our study) and it identifies and extracts all children of this mother node and uses their POS as features.

4.2 Corpus

The annotated data that we used for our experiments is the current version of the Berkeley FrameNet (Baker and Lowe, 1998) for English which consists of 139,437 annotated examples from the British National Corpus for 10,196 predicates. Among the predicates that FrameNet involvs, namely verbs, nouns, adjectives, and prepositions, we only considered *verbs*; as a result the data reduced to 61,792 annotated examples for 2,770 unique verb-frames.

In the next step, we removed all verbs that have only one frame as they are not ambiguous. Having only ambiguous verbs, the number of predicates reduced to 451 unique verbs. Out of these targets, there are only 37 verbs which have more than 100 annotated samples. Among these verbs, we concentrated on 14 verbs selected randomly; however, in the selection we tried to have a balance distribution of frames that the targets have. Therefore, we selected 4 targets (*phone*, *rush*, *scream*, *throw*) with

[1] http://maxent.sourceforge.net/

Table 1: Data distribution of the targets

Verb	Frames	Freq.	S	E	T
Bend	4	115	11	82	22
Feel	5	134	13	95	26
Follow	3	113	10	81	22
Forget	3	101	9	72	20
Hit	4	142	12	102	28
Look	3	183	15	134	34
Phone	2	166	14	121	31
Rise	4	110	11	77	22
Rush	2	168	14	123	31
Scream	2	148	12	108	28
Shake	4	104	10	73	21
Smell	3	146	13	106	27
Strike	3	105	10	75	20
Throw	2	155	13	113	29

two frames, 5 targets (*follow*, *forget*, *look*, *smell*, *strike*) with three frames, 4 targets (*bend*, *hit*, *rise*, *shake*) with four frames, and 1 target (*feel*) with five frames.

4.3 Data Distribution

The total amount of data prepared for the 14 verbs are divided into three non-overlapping sets in a balanced form in terms of both the number of the target predicate frames, and the relevant instances of each frame. In other words, the distribution should be such that different frames of the target verb is found in each of the three data sets. 10% is considered as initial seed data (S); 20% as test data (T), and the rest of 70% as extra unlabeled data (E). Table 1 presents the data distribution in which 5-fold cross-validation is performed to minimize the overfitting problem.

As mentioned, our proposed stopping criterion has two parameters, n and m, that should be tuned. For this purpose, we divided the 14 targets into the held-out set and the test set. To this aim, 7 targets, namely *feel*, *look*, *phone*, *rise*, *shake*, *smell*, and *throw* are selected as the held-out set; and 7 targets, namely *bend*, *follow*, *forget*, *hit*, *rush*, *scream*, and *strike* are used as the test set.

4.4 Results

Figures 2 and 3 illustrate the learning curves of the active learning process and random sampling as the baseline for the targets *look* and *rise*. The curves are the average of the 5 folds. As can be seen, in these targets our classifier has beaten the majority

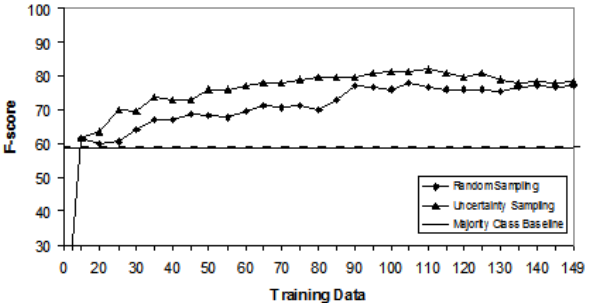

Figure 2: Learning curve of the verb *look* for 5 folds

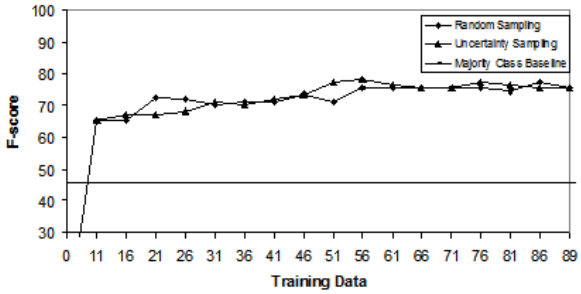

Figure 3: Learning curve of the verb *rise* for 5 folds

class baseline; and also active learning with uncertainty sampling has a relatively better performance than random sampling.

Figures 4 and 5 present the average variance curves of 5 folds for the two targets. These curves verify our assumption about the behavior of the variance curve as described in Section 3.2. As the graphs show, following our assumption the variability around the mean is tight in the early stages of training; then as the classifier is trained with more data, the variability around the mean spreads out; and finally, the variability will be tight again around the mean.

Applying our proposed stopping criterion, in each iteration we compute the variance of the classifier's confidence score for the selected samples in each fold. To evaluate how well our stopping criterion is, we have compared our results with the maximum average performance of the classifier for the 5 folds in which the whole data is labeled.

Applying our model on the held-out set, we found that n=2 is the best value based on our data set, so that we stop active learning when variance decreases in two sequential iterations; i.e.

$$V_i < V_{i-1} \text{ and } V_{i-1} < V_{i-2}.$$

Our idea is shown in Figure 6 for fold 5 of the target *rise*, such that the proposed stopping criterion is satisfied in iteration 11. As shown, the decrement of

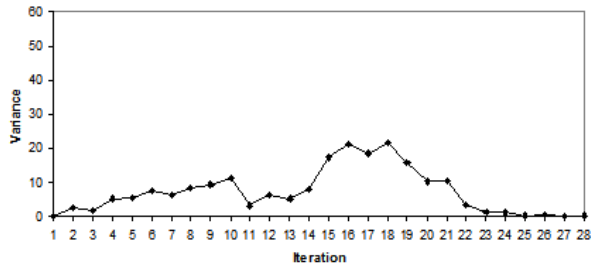

Figure 4: Variance curve of the verb *look* for 5 folds

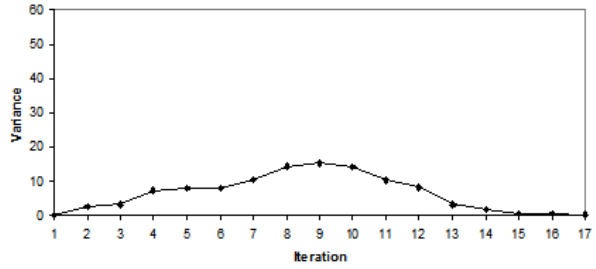

Figure 5: Variance curve of the verb *rise* for 5 folds

variance in iterations 3, 5, and 7 is the local maxima so that active learning does not stop in these iterations and they are ignored.

The summary of the result for the uncertainty sampling method of the test set is shown in Table 2 in which the *F*-score serves as the evaluation metric. Comparing the applied variance model as the stopping criterion on the test set with the maximum performance (M) of the uncertainty sampling as an upper bound in our experiment, we see that for two targets (*bend*, *rush*) the maximum performance of the classifier is achieved at the stopping point; for two targets (*follow*, *hit*) there is a minor reduction in the performance; while for the other targets (*forget*, *scream*, *strike*) there is a big loss in the performance. Averagely, the variance model achieved 92.66% of the maximum performance.

To determine the advantage of our stopping criterion, we present the total numbers of annotated instances (A) for each target, their relevant numbers of annotated instances for the maximum performance, and the variance model in Table 3. Av-

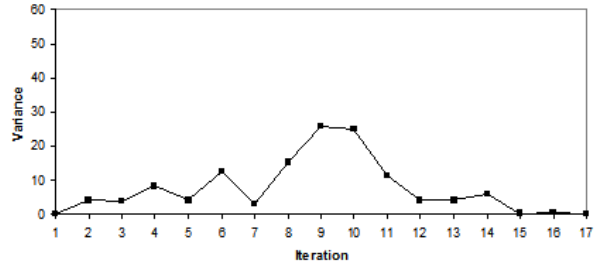

Figure 6: Variance curve of the verb *rise*

Table 2: The comparison of the average performance of the classifier (*F*-score) on the stopping point with the maximum performance in uncertainty sampling

Verb	M	VM
Bend	53.00	53.00
Follow	71.81	70.00
Forget	51.00	41.00
Hit	65.71	63.56
Rush	89.03	89.03
Scream	72.14	62.85
Strike	64.00	53.00
Average	**66.67**	**61.78**

Table 3: The comparison of the number of the annotated data for all data, at the maximum performance, and at the stopping point

Verb	A	M	VM
Bend	93	46	55
Follow	91	75	54
Forget	81	79	51
Hit	114	67	71
Rush	137	24	51
Scream	120	62	64
Strike	85	85	41
Average	**103**	**62.57**	**55.29**

eragely, if we have 103 samples for annotation, we need to annotate almost 63 instances to reach the maximum performance of 66.67%; while by applying our stopping criterion, the learning process stops by annotating at least 55 instances with 61.78% performance. I.e., annotating a smaller number of instances, our active learner achieves a near-optimum performance. It is worth to mention that since it is very difficult to achieve the upper bound of the classifier's performance automatically, all data is labeled to find the maximum performance of the classifier.

Looking carefully on the variance curves of the 5 folds of the held-out set, we have seen that in some iterations the decreased variance in two sequential iterations is very small and it may still stick in the local maxima as can be seen in iteration 8 of fold 3 of the target *look* in Figure 7.

To avoid sticking in such local maxima, we used the extended version of our original model and set a threshold (m) in the held-out set. Experimentally we found out that the decreasing variance in two sequential iterations must be bigger than 0.5; i.e.

$$V_i < V_{i-1} - 0.5 \text{ and } V_{i-1} < V_{i-2} - 0.5;$$

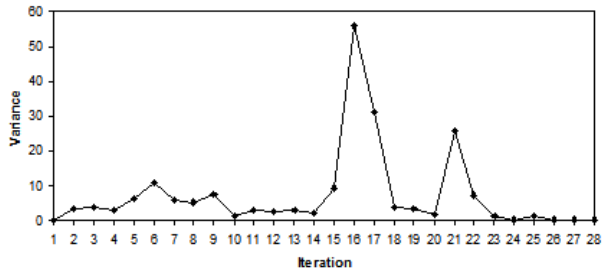

Figure 7: Variance curve of the verb *look*

so that in Figure 7 we stop in iteration 18. We applied the extended variance model on the test set and compared the results to our original variance model. We found out for two targets (*forget*, *scream*) the extended model has achieved a very good performance; for four targets (*follow*, *hit*, *rush*, *strike*) it was ineffective; and for one target (*bend*) it caused to have a small reduction in the performance.

The summary of the classifier performance after applying the extended model for uncertainty sampling is shown in Table 4. To ease the comparison, the performance of our original model is repeated in this table. As presented in the table, the average performance in the extended model has a 13.70% relative improvement compared to the average performance in the original variance model.

Table 4: The comparison of the average performance of the classifier (F-score) on the variance model and the extended variance model

Verb	**VM**	**EM**
Bend	53.00	52.00
Follow	70.00	70.00
Forget	41.00	46.00
Hit	63.56	63.56
Rush	89.03	89.03
Scream	62.85	63.56
Strike	53.00	53.00
Average	**61.78**	**62.45**

5 Related Work on Stopping Criteria

The simplest stopping criterion for active learning is when the training set has reached a desirable size or a predefined threshold. In this criterion, the active learning process repeatedly provides informative examples to the oracle for labeling, and updates the training set, until the desired size is obtained or the predefined stopping criterion is met. Practically, it is not clear how much annotation is sufficient for inducing a classifier with maximum effectiveness (Lewis and Gale, 1994).

Schohn and Cohn (2000) have used support vector machines (SVM) for document classification using the selective sampling method and they have proposed a criterion to stop the learning process in their task. Based on their idea, when there is no informative instance in the pool which is closer to the separating hyperplane than any of the support vectors, the margin exhausts and the learning process stops.

Zhu and Hovey (2007) have used a confidence-based approach for the stopping criteria by utlizing the *maximum confidence* and the *minimum error* of the classifier. The maximum confidence is based on the uncertainty measurement when the entropy of the selected unlabeled sample is less than a predefined threshold close to zero. The minimum error is the feedback from the oracle when active learning asks for the true label of the selected unlabeled sample and the accuracy prediction of the classifier for the selected unlabeled sample is larger than a predefined accuracy threshold. These criteria are considered as upper-bound and lower-bound of the stopping condition.

Zhu et al. (2008) proposed another stopping criterion based on a statistical learning approach called *minimum expected error strategy*. In this approach, the maximum effectiveness of the classifier is reached when the classifier's expected errors on future unlabeled data is minimum.

Vlachos (2008) has used the classifier confidence score as a stopping criterion for the uncertainty sampling. He has applied his model to two NLP tasks: text classification and named entity recognition. He has built his models with the SVM and the maximum entropy. The idea is when the confidence of the classifier remains at the same level or drops for a number of consecutive iterations, the learning process should terminate.

Laws and Schütze (2008) suggested three criteria -*minimal absolute performance*, *maximum possible performance*, and *convergence*- to stop active learning for name entity recognition using the SVM model with the uncertainty sampling method. In minimal absolute performance, a threshold is predefined by the user; and then the classifier estimates its own performance by using only the unlabeled reference test set. Since there is no available

labeled test set, the evaluation performance is not possible. The maximum possible performance is a confidence-based stopping criterion in which active learning is stopped where the optimal performance of the classifier is achieved. Again, in this approach there is no labeled test data to evaluate the performance. The convergence criterion is met when more examples from the pool of unlabeled data do not contribute more information to the classifier's performance, so that the classifier has reached its maximum performance. Laws and Schütze computed the convergence as the gradient of the classifier's estimated performance or uncertainty.

Tomanek and Hahn (2008) proposed a stopping criterion based on the performance of the classifier without requiring a labeled gold standard for a committee-based active learning on the name entity recognition application. In their criterion, they approximated the progression of the learning curve based on the disagreement among the committee members. They have used the *validation set agreement* curve as an adequate approximation for the progression of the learning curve. This curve was based on the data in each active learning iteration that makes the agreement values comparable between different active learning iterations.

Bloodgood and Vijay-Shanker (2009) explained three areas of stopping active learning that should be improved: applicability (restricting the usage in certain situation), lack of aggressive stopping (finding the stopping points which are too far, so more examples than necessary are annotated), instability (well working of a method on some data set but not the other data set). Further, they presented a stopping criterion based on *stabilizing predictions* that addresses each of the above three areas and provides a user-adjustable stopping behavior. In this method, the prediction of active learning was tested on examples which do not have to be labeled and it is stopped when the predictions are stabilized. This criterion was applied to text classification and named entity recognition tasks using the SVM and the maximum entropy models.

6 Summary and Future Work

In this paper, after a brief overview of frame semantics and active learning scenarios and query methods, we performed the frame assignment in the pool-based active learning with the uncertainty sampling method. To this end, we chose 14 frequent targets from FrameNet data set for our task.

One of the properties of active learning is its iterativness which should be stopped when the classifier has reached its maximum performance. Reaching this point is very difficult; therefore, we proposed a stopping criterion which stops active learning in a near-optimum point. This stopping criterion is based on the confidence score of the classifier on the extra unlabeled data such that it uses the variance of the classifier's confidence score for the predicted labels of a certain number of samples selected in each iteration. The advantage of this criterion is that there is no need to the labeled gold standard data and testing the performance of the classifier in each iteration. Based on this idea, we proposed a model which is satisfied by n sequential decrease on a variance curve. The original model is expanded by setting a threshold m on the amount of the decrement of variance in n sequential iterations. We believe that our proposed criterion can be applied to any active learning setting based on uncertainty sampling and it is not limited to the frame assignment.

To find out how effective our model is, we compared the achieved results of our variance model with the maximum performance of the classifier and we found that 92.66% of the performance is kept in the test data. In the extended variance model, we achieved a higher performance of the classifier in which 93.67% of the performance is kept.

For the future word, while in our current research the learner selects 5 instances in each iteration, this number could be different and investigation is needed to find out how much our proposed criterion depends on the K. The other possibility to expand our proposed model is using the variance of the classifier's confidence score for the predicted labels of the whole unlabeled data in each iteration and not the selected samples.

7 Acknowledgments

The author's special gratitude goes to Caroline Sporleder and Ines Rehbein at Saarland University for their support and helpful comments in the research. Masood Ghayoomi is funded by the German research council DFG under the contract number MU 2822/3-1.

References

C. F. Baker and C. J. Fillmore J. B. Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL*, pages 86–90, Montreal, QC.

J. Baldridge and M. Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of EMNLP*, pages 9–16, Barcelona, Spain.

M. Bloodgood and K. Vijay-Sarkar. 2009. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *13th Conf. on Computational Natural Language Learning*, pages 39–47, Boulder, Colorado.

B. Busser and R. Morante. 2005. Designing an active learning based system for corpus annotation. In *Revista de Procesamiento del Lenguaje Natural*, number 35, pages 375–381.

D. Cohn, A. L. Atlas, and R. E. Ladner. 1994. Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

K. Erk and S. Pado. 2006. Shalmaneser - a toolchain for shallow semantic parsing. In *Proceedings of LREC*, Genoa, Italy.

C. J. Fillmore. 1968. The case for case. In Emmon W. Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88, New York. Rinehart and Winston.

G. Haffari and A. Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the 47th ACL-IJCNLP*, Singapore.

F. Laws and H. Schütze. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd CoLing*, pages 465–472, Manchester.

D.D. Lewis and W. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR Conf. on Research and Development in IR*, pages 3–12.

I. Rehbein, J. Ruppenhofer, and J. Sunde. 2009. Majo - a toolkit for supervised word sense disambiguation and active learning. In *Proceedings of the 8th Int. Workshop on Treebanks and Linguistic Theories*, Milan, Italy.

G. Schohn and D. Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of 17th Int. Conf. on Machine Learning*, Stanford University.

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

C. A. Thompson, M.E. Califf, and R.J. Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the 16th Int. Conf. on Machine Learning*, pages 406–414.

K. Tomanek and U. Hahn. 2008. Approximating learning curves for active-learning-driven annotation. In *6th Int. Language Resources and Evaluation Conference*, pages 1319–1324.

A. Vlachos. 2008. A stopping criterion for active learning. *Journal of Computer, Speech and Language*, 22(3):295–312.

J. Zhu and E. Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the EMNLP-CoNLL*, pages 783–790, Prague.

J. Zhu, H. Wang, and E. Hovy. 2008. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the 3rd IJNLP*, pages 366–372, Heydarabad, India.