Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling

Dmitriy Dligach
Department of Computer Science
University of Colorado at Boulder
Dmitriy.Dligach@colorado.edu

Martha Palmer
Department of Linguistics
University of Colorado at Boulder
Martha.Palmer@colorado.edu

Abstract

Active Learning (AL) is typically initialized with a small seed of examples selected randomly. However, when the distribution of classes in the data is skewed, some classes may be missed, resulting in a slow learning progress. Our contribution is twofold: (1) we show that an unsupervised language modeling based technique is effective in selecting rare class examples, and (2) we use this technique for seeding AL and demonstrate that it leads to a higher learning rate. The evaluation is conducted in the context of word sense disambiguation.

1 Introduction

Active learning (AL) (Settles, 2009) has become a popular research field due to its potential benefits: it can lead to drastic reductions in the amount of annotation that is necessary for training a highly accurate statistical classifier. Unlike in a random sampling approach, where unlabeled data is selected for annotation randomly, AL delegates the selection of unlabeled data to the classifier. In a typical AL setup, a classifier is trained on a small sample of the data (usually selected randomly), known as the seed examples. The classifier is subsequently applied to a pool of unlabeled data with the purpose of selecting additional examples that the classifier views as informative. The selected data is annotated and the cycle is repeated, allowing the learner to quickly refine the decision boundary between the classes.

Unfortunately, AL is susceptible to a shortcoming known as the *missed cluster effect* (Schütze et al., 2006) and its special case called the *missed class effect* (Tomanek et al., 2009). The missed cluster effect is a consequence of the fact that seed examples influence the direction the learner takes in its exploration of the instance space. Whenever the seed does not contain the examples of a certain cluster that is representative of a group of examples in the data, the learner may become overconfident about the class membership of this cluster (particularly if it lies far from the decision boundary). As a result, the learner spends a lot of time exploring one region of the instance space at the expense of missing another. This problem can become especially severe, when the class distribution in the data is skewed: a randomly selected seed may not adequately represent all the classes or even miss certain classes altogether. Consider a binary classification task where rare class examples constitute 5% of the data (a frequent scenario in e.g. word sense disambiguation). If 10 examples are chosen randomly for seeding AL, the probability that *none* of the rare class examples will make it to the seed is 60% [1]. Thus, there is a high probability that AL would stall, selecting only the examples of the predominant class over the course of many iterations. At the same time, if we had a way to ensure that examples of the rare class were present in the seed, AL would be able to select the examples of both classes, efficiently clarifying the decision boundary and ultimately producing an accurate classifier.

Tomanek et al. (2009) simulated these scenarios using *manually* constructed seed sets. They demonstrated that seeding AL with a data set that is artificially enriched with rare class examples indeed leads to a higher learning rate comparing to randomly

[1] Calculated using Binomial distribution

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 6–10,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

sampled and predominant class enriched seeds. In this paper, we propose a simple *automatic* approach for selecting the seeds that are rich in the examples of the rare class. We then demonstrate that this approach to seed selection accelerates AL. Finally, we analyze the mechanism of this acceleration.

2 Approach

Language Model (LM) Sampling is a simple unsupervised technique for selecting unlabeled data that is enriched with rare class examples. LM sampling involves training a LM on a corpus of unlabeled candidate examples and selecting the examples with low LM probability. Dligach and Palmer (2009) used this technique in the context of word sense disambiguation and showed that rare sense examples tend to concentrate among the examples with low probability. Unfortunately these authors provided a limited evaluation of this technique: they looked at its effectiveness only at a single selection size. We provide a more convincing evaluation in which the effectiveness of this approach is examined for all sizes of the selected data.

Seed Selection for AL is typically done randomly. However, for datasets with a skewed distribution of classes, rare class examples may end up being underrepresented. We propose to use LM sampling for seed selection, which captures more examples of rare classes than random selection, thus leading to a faster learning progress.

3 Evaluation

3.1 Data

For our evaluation, we needed a dataset that is characterized by a skewed class distribution. This phenomenon is pervasive in word sense data. A large word sense annotated corpus has recently been released by the OntoNotes (Hovy et al., 2006; Weischedel et al., 2009) project. For clarity of evaluation, we identify a set of verbs that satisfy three criteria: (1) the number of senses is two, (2) the number of annotated examples is at least 100, (3) the proportion of the rare sense is at most 20%. The following 25 verbs satisfy these criteria: *account, add, admit, allow, announce, approve, compare, demand, exist, expand, expect, explain, focus, include, invest, issue, point, promote, protect, receive, remain, replace, strengthen, wait, wonder*. The average number of examples for these verbs is 232. In supervised word sense disambiguation, a single model per word is typically trained and that is the approach we take. Thus, we conduct our evaluation using 25 different data sets. We report the averages across these 25 data sets. In our evaluation, we use a state-of-the-art word sense disambiguation system (Dligach and Palmer, 2008), that utilizes rich linguistic features to capture the contexts of ambiguous words.

3.2 Rare Sense Retrieval

The success of our approach to seeding AL hinges on the ability of LM sampling to discover rare class examples better than random sampling. In this experiment, we demonstrate that LM sampling outperforms random sampling for every selection size. For each verb we conduct an experiment in which we select the instances of this verb using both methods. We measure the *recall* of the rare sense, which we calculate as the ratio of the number of selected rare sense examples to the total number of rare sense examples for this verb.

We train a LM (Stolcke, 2002) on the corpora from which OntoNotes data originates: the Wall Street Journal, English Broadcast News, English Conversation, and the Brown corpus. For each verb, we compute the LM probability for each instance of this verb and sort the instances by probability. In the course of the experiment, we select one example with the smallest probability and move it to the set of selected examples. We then measure the recall of the rare sense for the selected examples. We continue in this fashion until all the examples have been selected. We use random sampling as a baseline, which is obtained by continuously selecting a single example randomly. We continue until all the examples have been selected. At the end of the experiment, we have produced two recall curves, which measure the recall of the rare sense retrieval for this verb at various sizes of selected data. Due to the lack of space, we do not show the plots that display these curves for individual verbs. Instead, in Figure 1 we display the curves that are averaged across all verbs. At every selection size, LM sampling results in a higher recall of the rare sense. The average difference across all selection sizes is 11%.

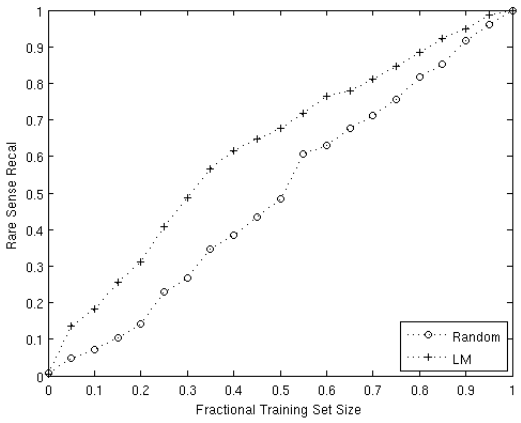

Figure 1: Average recall of rare sense retrieval for LM and random sampling by relative size of training set

3.3 Classic and Selectively Seeded AL

In this experiment, we seed AL using LM sampling and compare how this selectively seeded AL performs in comparison with classic (randomly-seeded) AL. Our experimental setup is typical for an active learning study. We split the set of annotated examples for a verb into 90% and 10% parts. The 90% part is used as a pool of unlabeled data. The 10% part is used as a test set. We begin classic AL by randomly selecting 10% of the examples from the pool to use as seeds. We train a maximum entropy model (Le, 2004) using these seeds. We then repeatedly apply the model to the remaining examples in the pool: on each iteration of AL, we draw a single most informative example from the pool. The informativeness is estimated using prediction margin (Schein and Ungar, 2007), which is computed as $|P(c_1|x) - P(c_2|x)|$, where c_1 and c_2 are the two most probable classes of example x according to the model. The selected example is moved to the training set. On each iteration, we also keep track of how accurately the current model classifies the held out test set.

In parallel, we conduct a selectively seeded AL experiment that is identical to the classic one but with one crucial difference: instead of selecting the seed examples randomly, we select them using LM sampling by identifying 10% of the examples from the pool with the smallest LM probability. We also produce a random sampling curve to be used as a baseline. At the end of this experiment we have obtained three learning curves: for classic AL, for selectively seeded AL, and for the random sampling baseline. The final learning curves for each verb are produced by averaging the learning curves from ten different trials.

Figure 2 presents the average accuracy of selectively seeded AL (top curve), classic AL (middle curve) and the random sampling baseline (bottom curve) at various fractions of the total size of the training set. The size of zero corresponds to a training set consisting only of the seed examples. The size of one corresponds to a training set consisting of all the examples in the pool labeled. The accuracy at a given size was averaged across all 25 verbs.

It is clear that LM-seeded AL accelerates learning: it reaches the same performance as classic AL with less training data. LM-seeded AL also reaches a higher classification accuracy (if stopped at its peak). We will analyze this somewhat surprising behavior in the next section. The difference between the classic and LM-seeded curves is statistically significant (p = 0.0174) [2].

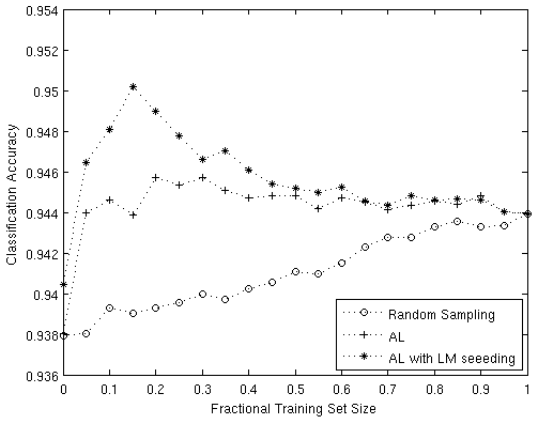

Figure 2: Randomly and LM-seeded AL. Random sampling baseline is also shown.

3.4 Why LM Seeding Produces Better Results

For random sampling, the system achieves its best accuracy, 94.4%, when the entire pool of unlabeled examples is labeled. The goal of a typical AL study is to demonstrate that the *same* accuracy can be

[2] We compute the average area under the curve for each type of AL and use Wilcoxon signed rank test to test whether the difference between the averages is significant.

achieved with less labeled data. For example, in our case, classic AL reaches the best random sampling accuracy with only about 5% of the data. However, it is interesting to notice that LM-seeded AL actually reaches a *higher* accuracy, 95%, during early stages of learning (at 15% of the total training set size). We believe this phenomenon takes place due to overfitting the predominant class: as the model receives new data (and therefore more and more examples of the predominant class), it begins to mislabel more and more examples of the rare class. A similar idea has been expressed in literature (Weiss, 1995; Kubat and Matwin, 1997; Japkowicz, 2001; Weiss, 2004; Chen et al., 2006), however it has never been verified in the context of AL.

To verify our hypothesis, we conduct an experiment. The experimental setup is the same as in section 3.3. However, instead of measuring the *accuracy* on the test set, we resort to different metrics that reflect how accurately the classifier labels the instances of the rare class in the held out test set. These metrics are the recall and precision for the rare class. *Recall* is the ratio of the correctly labeled examples of the rare class and the total number of instances of the rare class. *Precision* is the ratio of the correctly labeled examples of the rare class and the number of instances labeled as that class. Results are in Figures 3 and 4.

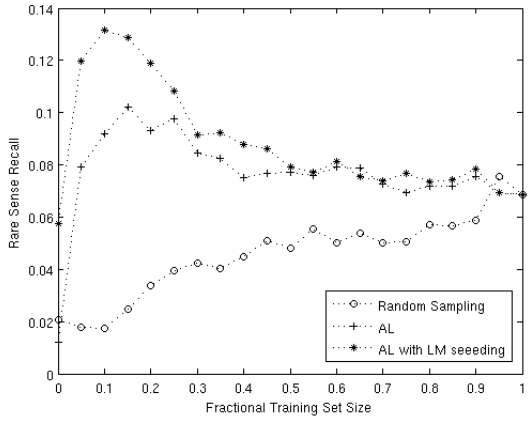

Figure 3: Rare sense classification recall

Observe that for LM-seeded AL, the recall peaks at first and begins to decline later. Thus the classifier makes progressively more errors on the rare class as more labeled examples are being received.

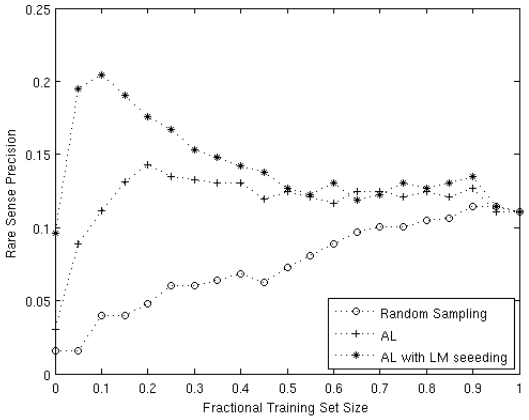

Figure 4: Rare sense classification precision

This is consistent with our hypothesis that the classifier overfits the predominant class. When all the data is labeled, the recall decreases from about 13% to only 7%, an almost 50% drop. The reason that the system achieved a higher level of recall at first is due to the fact that AL was seeded with LM selected data, which has a higher content of rare classes (as we demonstrated in the first experiment). The availability of the extra examples of the rare class allows the classifier to label the instances of this class in the test set more accurately, which in turn boosts the overall accuracy.

4 Conclusion and Future Work

We introduced a novel approach to seeding AL, in which the seeds are selected from the examples with low LM probability. This approach selects more rare class examples than random sampling, resulting in more rapid learning and, more importantly, leading to a classifier that performs better on rare class examples. As a consequence of this, the overall classification accuracy is higher than that for classic AL.

Our plans for future work include improving our LM by incorporating syntactic information such as POS tags. This should result in better performance on the rare classes, which is currently still low. We also plan to experiment with other unsupervised techniques, such as clustering and outlier detection, that can lead to better retrieval of rare classes. Finally, we plan to investigate the applicability of our approach to a multi-class scenario.

Acknowledgements

We gratefully acknowledge the support of the National Science Foundation Grant NSF-0715078, Consistent Criteria for Word Sense Disambiguation, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, a subcontract from the BBN-AGILE Team. 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.

References

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

Dmitriy Dligach and Martha Palmer. 2008. Novel semantic features for verb sense disambiguation. In *HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 29–32, Morristown, NJ, USA. Association for Computational Linguistics.

Dmitriy Dligach and Martha. Palmer. 2009. Using language modeling to select useful annotation data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 25–30. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

Nathalie Japkowicz. 2001. Concept-learning in the presence of between-class and within-class imbalances. In *AI '01: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, pages 67–77, London, UK. Springer-Verlag.

M. Kubat and S. Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Citeseer.

Zhang Le, 2004. *Maximum Entropy Modeling Toolkit for Python and C++*.

A.I. Schein and L.H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265.

H. Schütze, E. Velipasaoglu, and J.O. Pedersen. 2006. Performance thresholding in practical text classification. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 662–671. ACM.

Burr Settles. 2009. Active learning literature survey. In *Computer Sciences Technical Report 1648 University of Wisconsin-Madison*.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing, Denver, Colorado.*, pages 901–904.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. 2009. On proper unit selection in active learning: co-selection effects for named entity recognition. In *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Morristown, NJ, USA. Association for Computational Linguistics.

R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R Belvin, S Pradan, L. Ramshaw, and N. Xue, 2009. *OntoNotes: A Large Training Corpus for Enhanced Processing*, chapter in Global Automatic Language Exploitation, pages 54–63. Springer Verglag.

G.M. Weiss. 1995. Learning with rare cases and small disjuncts. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 558–565. Citeseer.

G.M. Weiss. 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19.