Active Learning for New Domains in Natural Language Understanding

Stanislav Peshterliev, John Kearney, Abhyuday Jagannatha, Imre Kiss, Spyros Matsoukas
Alexa Machine Learning, Amazon.com
{stanislp,jkearn,abhyudj,ikiss,matsouka}@amazon.com

Abstract

We explore active learning (AL) for improving the accuracy of new domains in a natural language understanding (NLU) system. We propose an algorithm called Majority-CRF that uses an ensemble of classification models to guide the selection of relevant utterances, as well as a sequence labeling model to help prioritize informative examples. Experiments with three domains show that Majority-CRF achieves 6.6%-9% relative error rate reduction compared to random sampling with the same annotation budget, and statistically significant improvements compared to other AL approaches. Additionally, case studies with human-in-the-loop AL on six new domains show 4.6%-9% improvement on an existing NLU system.

1 Introduction

Intelligent voice assistants (IVA) such as Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana, are becoming increasingly popular. For IVA, natural language understanding (NLU) is a main component (De Mori et al., 2008), in conjunction with automatic speech recognition (ASR) and dialog management (DM). ASR converts user's speech to text. Then, the text is passed to NLU for classifying the action or "intent" that the user wants to invoke (e.g., PlayMusicIntent, TurnOnIntent, BuyItemIntent) and recognizing named-entities (e.g., Artist, Genre, City). Based on the NLU output, DM decides the appropriate response, which could be starting a song playback or turning off lights. NLU systems for IVA support functionality in a wide range of domains, such as music, weather, and traffic. Also, an important requirement is the ability to add support for new domains.

The NLU models for Intent Classification (IC) and Named Entity Recognition (NER) use machine learning to recognize variation in natural language. Diverse, annotated training data collected from IVA users, or "annotated live utterances," are essential for these models to achieve good performance. As such, new domains frequently exhibit suboptimal performance due to a lack of annotated live utterances. While an initial training dataset can be bootstrapped using grammar generated utterances and crowdsourced collection (Amazon Mechanical Turk), the performance that can be achieved using these approaches is limited because of the unexpected discrepancies between anticipated and live usage. Thus, a mechanism is required to select live utterances to be manually annotated for enriching the training dataset.

Random sampling is a common method for selecting live utterances for annotation. However, in an IVA setting with many users, the number of available live utterances is vast. Meanwhile, due to the high cost of manual annotation, only a small percentage of utterances can be annotated. As such, in a random sample of live data, the number of utterances relevant to new domains may be small. Moreover, those utterances may not be informative, where informative utterances are those that, if annotated and added to the training data, reduce the error rates of the NLU system. Thus, for new domains, we want a sampling procedure which selects utterances that are both relevant and informative.

Active learning (AL) (Settles, 2009) refers to machine learning methods that can interact with the sampling procedure and guide the selection of data for annotation. In this work, we explore using AL for live utterance selection for new domains in NLU. Authors have successfully applied AL techniques to NLU systems with little annotated data overall (Tur et al., 2003; Shen et al., 2004). The difference with our work is that, to the best of our knowledge, there is little published AL research that focuses on data selection explicitly targeting new domains.

Proceedings of NAACL-HLT 2019, pages 90–96
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

We compare the efficacy of least-confidence (Lewis and Catlett, 1994) and query-by-committee (Freund et al., 1997) AL for new domains. Moreover, we propose an AL algorithm called Majority-CRF, designed to improve both IC and NER of an NLU system. Majority-CRF uses an ensemble of classification models to guide the selection of relevant utterances, as well as a sequence labeling model to help prioritize informative examples. Simulation experiments on three different new domains show that Majority-CRF achieves 6.6%-9% relative improvements in-domain compared to random sampling, as well as significant improvements compared to other active learning approaches.

2 Related Work

Expected model change (Settles et al., 2008) and expected error reduction (Roy and McCallum, 2001) are AL approaches based on decision theory. Expected model change tries to select utterances that cause the greatest change on the model. Similarly, expected error reduction tries to select utterances that are going to maximally reduce generalization error. Both methods provide sophisticated ways for ascertaining the value of annotating an utterance. However, they require computing an expectation across all possible ways to label the utterance, which is computationally expensive for NER and IC models with many labels and millions of parameters. Instead, approaches to AL for NLU generally require finding a proxy, such as model uncertainty, to estimate the value of getting specific points annotated.

Tur et al. studied least-confidence and query-by-committee disagreement AL approaches for reducing the annotation effort (Tur et al., 2005, 2003). Both performed better than random sampling, and the authors concluded that the overall annotation effort could be halved. We investigate both of these approaches, but also a variety of new algorithms that build upon these basic ideas.

Schutze et al. (Schütze et al., 2006) showed that AL is susceptible to the missed cluster effect when selection focuses only on low confidence examples around the existing decision boundary, missing important clusters of data that receive high confidence. They conclude that AL may produce a sub-optimal classifier compared to random sampling with a large budget. To solve this problem Osugi et al. (Osugi et al., 2005) proposed an AL algorithm that can balance exploitation (sampling around the decision boundary) and exploration (random sampling) by reallocating the sampling budget between the two. In our setting, we start with a representative seed dataset, then we iteratively select and annotate small batches of data that are used as feedback in subsequent selections, such that extensive exploration is not required.

To improve AL, Hong-Kwang and Vaibhava (Kuo and Goel, 2005) proposed to exploit the similarity between instances. Their results show improvements over simple confidence-based selection for data sizes of less than 5,000 utterances. A computational limitation of the approach is that it requires computing the pairwise utterance similarity, an $\mathcal{O}(N^2)$ operation that is slow for millions of utterances available in production IVA. However, their approach could be potentially sped-up with techniques like locality-sensitive hashing.

3 Active Learning For New Domains

We first discuss random sampling baselines and standard active learning approaches. Then, we describe the Majority-CRF algorithm and the other AL algorithms that we tested.

3.1 Random Sampling Baselines

A common strategy to select live utterances for annotation is random sampling. We consider two baselines: uniform random sampling and domain random sampling.

Uniform random sampling is widespread because it provides unbiased samples of the live utterance distribution. However, the samples contain fewer utterances for new domains because of their low usage frequency. Thus, under a limited annotation budget, accuracy improvements on new domains are limited.

Domain random sampling uses the predicted NLU domain to provide samples of live utterances more relevant to the target domains. However, this approach does not select the most informative utterances.

3.2 Active Learning Baselines

AL algorithms can select relevant and informative utterances for annotation. Two popular AL approaches are least-confidence and query-by-committee.

Least-confidence (Lewis and Catlett, 1994) involves processing live data with the NLU models and prioritizing selection of the utterances with the least confidence. The intuition is that utterances with low confidence are difficult, and "teaching" the models how they should be labeled is informative. However, a weakness of this method is that out-of-domain or irrelevant utterances are likely to be selected due to low confidence. This weakness can be alleviated by looking at instances with medium confidence using measures such as least margin between the top-n hypotheses (Scheffer et al., 2001) or highest Shannon entropy (Settles and Craven, 2008).

Query-by-committee (QBC) (Freund et al., 1997) uses different classifiers (e.g., SVMs, MaxEnt, Random Forests) that are trained on the existing annotated data. Each classifier is applied independently to every candidate and the utterances assigned the most diverse labels are prioritized for annotation. One problem with this approach is that, depending on the model and the size of the committee, it could be computationally expensive to apply on large datasets.

3.3 Majority-CRF Algorithm

Majority-CRF is a confidence-based AL algorithm that uses models trained on the available NLU training set but does not rely on predictions from the full NLU system. Its simplicity compared to a full NLU system offers several advantages. First, fast incremental training with the selected annotated data. Second, fast predictions on millions of utterances. Third, the selected data is not biased to the current NLU models, which makes our approach reusable even if the models change.

Algorithm 1 shows a generic AL procedure that we use to implement Majority-CRF, as well as other AL algorithms that we tested. We train an ensemble of models on positive data from the target domain of interest (e.g., Books) and negative data that is everything not in the target domain (e.g., Music, Videos). Then, we use the models to filter and prioritize a batch of utterances for annotation. After the batch is annotated, we retrain the models with the new data and repeat the process.

To alleviate the tendency of the least-confidence approaches to select irrelevant data, we add unsupported utterances and sentence fragments to the negative class training data of the AL models. This helps keep noisy utterances on the negative side of the decision boundary, so that they can be eliminated during filtering. Note that, when targeting several domains at a time, we run the selection procedure independently and then deduplicate the utterances before sending them for annotation.

Algorithm 1 Generic AL procedure that selects data for a target domain

```
Inputs:
    $D \leftarrow$ positive and negative training data
    $P \leftarrow$ pool of unannotated live utterances
    $i \leftarrow$ iterations, $m \leftarrow$ mini-batch size
Parameters:
    $\{\mathcal{M}^k\} \leftarrow$ set of selection models
    $\mathcal{F} \leftarrow$ filtering function
    $\mathcal{S} \leftarrow$ scoring function
Procedure:
repeat $i$ iterations
    Train selection models $\{\mathcal{M}^k\}$ on $D$
    $\forall\ x_i \in P$ obtain prediction scores $y_i^k = \mathcal{M}^k(x_i)$
    $P' \leftarrow \{x_i \in P\ :\ \mathcal{F}(y_i^0..y_i^k)\ \}$
    $C \leftarrow \{x_i \in P'\ :\ m$ with the smallest score $\mathcal{S}(y_i^0..y_i^k)\}$
    Send $C$ for manual annotation
    After annotation is done $D \leftarrow D \cup C$ and $P \leftarrow P \setminus C$
until
```

Models. We experimented with n-gram linear binary classifiers trained to minimize different loss functions: $\mathcal{M}^{lg} \leftarrow$ logistic, $\mathcal{M}^{hg} \leftarrow$ hinge, and $\mathcal{M}^{sq} \leftarrow$ squared. Each classifier is trained to distinguish between positive and negative data and learns a different decision boundary. Note that we use the raw unnormalized prediction scores $\{y_i^{lg}, y_i^{hg}, y_i^{sq}\}$ (no sigmoid applied) that can be interpreted as distances between the utterance x_i and the classifiers decision boundaries at $y = 0$. The classifiers are implemented in Vowpal Wabbit (Langford et al., 2007) with $\{1, 2, 3\}$-gram features. To directly target the NER task, we used an additional $\mathcal{M}^{cf} \leftarrow$ CRF, trained on the NER labels of the target domain.

Filtering function. We experimented with $\mathcal{F}^{maj} \leftarrow \sum \text{sgn}(y^k) > 0$, i.e., keep only majority positive prediction from the binary classifiers, and $\mathcal{F}^{dis} \leftarrow \sum \text{sgn}(y^k) \in \{-1, 1\}$, i.e., keep only prediction where there is at least one disagreement.

Scoring function. When the set of models $\{\mathcal{M}^k\}$ consists of only binary classifiers, we combine the classifier scores using either the sum of

Algorithm	Models $\{\mathcal{M}^i\}$	Filter $\mathcal{F}$	Scoring $\mathcal{S}$
AL-Logistic	lg	$\text{sgn}(y^{lg}) > 0$	y^{lg}
QBC-SA	lg, sq, hg	$\sum \text{sgn}(y^k) \in \{-1, 1\}$	$\sum \lvert y^k \rvert$
QBC-AS	lg, sq, hg	$\sum \text{sgn}(y^k) \in \{-1, 1\}$	$\lvert \sum y^k \rvert$
Majority-SA	lg, sq, hg	$\sum \text{sgn}(y^k) > 0$	$\sum \lvert y^k \rvert$
Majority-AS	lg, sq, hg	$\sum \text{sgn}(y^k) > 0$	$\lvert \sum y^k \rvert$
QBC-CRF	lg, sq, hg, CRF	$\sum \text{sgn}(y^k) \in \{-1, 1\}$	$p^{lg} \times p^{crf}$
Majority-CRF	lg, sq, hg, CRF	$\sum \text{sgn}(y^k) > 0$	$p^{lg} \times p^{crf}$

Table 1: AL algorithms evaluated. lg, sq, hg refer to binary classifiers (committee members) trained with logistic, squared and hinge loss functions, respectively. y^i denotes the score of committee member i, p^{crf} denotes the confidence of the CRF model and $p^{lg} = (1 + e^{-y^{lg}})^{-1}$ denotes the confidence of the logistic classifier. In all cases, we prioritize by smallest score $\mathcal{S}$.

absolutes $\mathcal{S}^{sa} \leftarrow \sum |y_i^k|$ or the absolute sum $\mathcal{S}^{as} \leftarrow |\sum y_i^k|$. $\mathcal{S}^{sa}$ prioritizes utterances where all scores are small (i.e., close to all decision boundaries), and $\mathcal{S}^{as}$ prioritizes utterances where either all scores are small or there is large disagreement between classifiers (e.g., one score is large negative, another is large positive, and the third is small). Both $\mathcal{S}^{sa}$ and $\mathcal{S}^{as}$ can be seen as generalization of least-confidence to a committee of classifiers. When the set of models $\{\mathcal{M}^k\}$ includes a CRF model $\mathcal{M}^{cf}$, we compute the score with $\mathcal{S}^{cg} \leftarrow P_{cf}(i) \times P_{lg}(i)$, i.e., the CRF probability $P_{cf}(i)$ multiplied by the logistic classifier probability $P_{lg}(i) = \sigma(y_i^{lg})$, where σ is the sigmoid function. Note that we ignore the outputs of the squared and hinge classifiers for scoring, though they are still be used for filtering.

The full set of configurations we evaluated is given in Table 1, which specifies the choice of parameters $\{\mathcal{M}^k\}, \mathcal{F}, \mathcal{S}$ used in Algorithm 1.

AL-Logistic and QBC serve as baseline AL algorithms. The QBC-CRF and Majority-CRF models combine the IC focused binary classifier scores with the NER focused sequence labeling scores and use filtering by disagreement and majority (respectively) to select informative utterances. To the best of our knowledge, this is a novel architecture for active learning in NLU.

Mamitsuka et al. (Mamitsuka et al., 1998) proposed bagging to build classifier committees for AL. Bagging refers to random sampling with replacement of the original training data to create diverse classifiers. We experimented with bagging but found that it is not better than using different classifiers.

4 Experimental Results

4.1 Evaluation Metrics

We use Slot Error Rate (SER) (Makhoul et al., 1999), including the intent as slot, to evaluate the overall predictive performance of the NLU models. SER as the ratio of the number of slot prediction errors to the total number of reference slots. Errors are insertions, substitutions and deletions. We treat the intent misclassifications as substitution errors.

4.2 Simulated Active Learning

AL requires manual annotations which are costly. Therefore, to conduct multiple controlled experiments with different selection algorithms, we simulated AL by taking a subset of the available annotated training data as the unannotated candidate pool, and “hiding” the annotations. As such, the NLU system and AL algorithm had a small pool of annotated utterances for simulated “new” domains. Then, the AL algorithm was allowed to choose relevant utterances from the simulated candidate pool. Once an utterance is selected, its annotation is revealed to the AL algorithm, as well as to the full NLU system.

Dataset. We conducted experiments using an internal test dataset of 750K randomly sampled live utterances, and a training dataset of 42M utterances containing a combination of grammar generated and randomly sampled live utterances. The dataset covers 24 domains, including Music, Shopping, Local Search, Sports, Books, Cinema and Calendar.

NLU System. Our NLU system has one set of IC and NER models per domain. The IC model predicts one of its in-domain intents or a special out-of-domain intent which helps with domain classification. The IC and NER predictions are ranked into a single n-best list based on model confidences (Su et al., 2018). We use MaxEnt (Berger et al., 1996) models for IC and the CRF models for NER (Lafferty et al., 2001).

Experimental Design. We split the training data into a 12M utterances initial training set for IC and NER, and a 30M utterance candidate pool for selection. We choose Books, Local Search, and Cinema as target domains to simulate the AL al-

Domain	Train	Test	Examples
Books	290K	13K	“search in mystery books” “read me a book”
Local Search	260K	16K	“mexican food nearby” “pick the top bank”
Cinema	270K	9K	“more about hulk” “what’s playing in theaters”

Table 2: Simulated ”new“ target domains for AL experiments. The target domain initial training datasets are 90% grammar generated data. The other 21 ”non-new“ domains have on average 550k initial training datasets with 60% grammar generated data and 40% live data.

gorithms, see Table 2. Each target domain had 550-650K utterances in the candidate pool. The rest of the 21 non-target domains have 28.5M utterances in the candidate pool. We also added 100K sentence fragments and out-of-domain utterances to the candidate pool, which allows us to compare the susceptibility of different algorithms to noisy or irrelevant data. This experimental setup attempts to simulate the production IVA use case where the candidate pool has a large proportion of utterances that belong to different domains.

We employed the different AL algorithms to select 12K utterances per domain from the candidate pool, for a total 36K utterance annotation budget. Also, we evaluated uniform (Rand-Uniform) and domain (Rand-Domain) random sampling with the same total budget. We ran each AL configuration twice and average the SER scores to account for fluctuations in selection caused by the stochasticity in model training. For random sampling, we ran each selection five times.

4.2.1 Simulated Active Learning Results

Table 3 shows the experimental results for the target domains Books, Local Search, and Cinema. For each experiment, we add all AL selected data (in- and out-of-domain), and evaluate SER for the full NLU system.

We test for statistically significant improvements using the Wilcoxon test (Hollander et al., 2013) with 1000 bootstrap resamples and p-value < 0.05.

Random Baselines. As expected, Rand-Uniform selected few relevant utterances for the target domains due to their low frequency in the candidate pool. Rand-Domain selects relevant utterances for the target domains, achieving statistically significant SER improvements compared to Rand-Uniform. However, the overall gains are small, around 1% relative per target domain. A significant factor for Rand-Domain’s limited improvement is that it tends to capture frequently-occurring utterances that the NLU models can already recognize without errors. As such, all AL configurations achieved statistically significant SER gains compared to the random baselines.

Single Model Algorithms. AL-Logistic, which carries out a single iteration of confidence-based selection, exhibits a statistically significant reduction in SER relative to Rand-Domain. Moreover, using six iterations (i.e., i=6) further reduced SER by a statistically significant 1%-2% relative to AL-Logistic(i=1), and resulted in the selection of 200 fewer unsupported utterances. This result demonstrates the importance of incremental selection for iteratively refining the selection model.

Committee Algorithms. AL algorithms incorporating a committee of models outperformed those based on single models by a statistically significant 1-2% ΔSER. The *majority* algorithms performed slightly better than the QBC algorithms and were able to collect more in-domain utterances. The absolute sum scoring function $\mathcal{S}^{as}$ performed slightly better than the sum of absolutes $\mathcal{S}^{sa}$ for both QBC and Majority. Amongst all committee algorithms, Majority-AS performed best, but the differences with the other committee algorithms are not statistically significant.

Committee and CRF Algorithms. AL algorithms incorporating a CRF model tended to outperform purely classification-based approaches, indicating the importance of specifically targeting the NER task. The Majority-CRF algorithm achieves a statistically significant SER improvement of 1-2% compared to Majority-AS (the best configuration without the CRF). Again, the disagreement-based QBC-CRF algorithm performed worse that the majority algorithm across target domains. This difference was statistically significant on Books, but not on Cinema and Local Search.

In summary, AL yields more rapid improvements not only by selecting utterances relevant to the target domain but also by trying to select the most informative utterances. For instance, although the AL algorithms selected 40-50% false positive utterances from non-target domains, whereas Rand-Domain selected only around 20% false positives, the AL algorithms still outperformed Rand-Domain. This indicates that labeling ambiguous false positives helps resolve existing confusions

Algorithm Group	Algorithm ($i = 6$)	Overall	Books		Local Search		Cinema		Non-Target
		#Utt	#Utt	ΔSER	#Utt	ΔSER	#Utt	ΔSER	#Utt
Random	Rand-Uniform	35.8K	747	1.20	672	3.37	547	0.57	33.8K
	Rand-Domain	35.7K	9853	1.52	9453	4.23	9541	1.75	06.8K
Single Model	AL-Logistic(i=1)	34.9K	5405	4.76	7092	6.54	5224	6.09	17.1K
	AL-Logistic	35.1k	5524	6.77	7709	7.24	5330	7.29	16.5K
Committee Models	QBC-AS	35.0K	4768	7.18	7869	8.57	4706	8.72	17.6K
	QBC-SA	35.0K	4705	7.12	7721	8.96	4790	7.52	17.7K
	Majority-AS	35.1K	5389	7.66	8013	9.07	5526	8.98	16.1K
	Majority-SA	35.1K	5267	7.35	8196	8.46	5193	8.42	16.4K
Committee and CRF	QBC-CRF	35.1K	3653	7.44	6593	9.78	4064	10.26	20.7K
	Majority-CRF	35.1K	6541	**8.42**	8552	**9.92**	6951	**11.05**	13.0K

Table 3: Simulation experimental results with 36K annotation budget. ΔSER is % relative reduction is SER compared to the initial model: Books SER 30.59, Local Search SER 39.09, Cinema SER 38.71. Higher ΔSER is better. The best result is in bold, and the second best is underlined. The $i = 1$ means selection in a single iteration, otherwise if not specified selection is in six iterations ($i = 6$). Overall #Utt shows the remaining from the 36K selected after removing the sentence fragments and out-of-domain utterances. Both target and non-target domains IC and NER models are re-retrained with the new data.

between domains. Another important observation is that majority filtering $\mathcal{F}^{maj}$ performs better than QBC disagreement filtering $\mathcal{F}^{dis}$ across all of our experiments. A possible reason for this is that majority filtering selects a better balance of boundary utterances for classification and in-domain utterances for NER. Finally, the Majority-CRF results show that incorporating the CRF model improves the performance of the committee algorithms. We assume this is because incorporation of a CRF-based confidence directly targets the NER task.

4.3 Human-in-the-loop Active Learning

We also performed AL for six new NLU domains with human-in-the-loop annotators and live user data. We used the Majority-SA configuration for simplicity in these case studies. We ran the AL selection for 5-10 iterations with varying batch sizes between 1000-2000.

Domain	ΔSER	#Utt Selected	#Utt Testset
Recipes	8.97	24.1K	4.7K
LiveTV	6.92	11.6K	1.8K
OpeningHours	7.05	6.8K	583
Navigation	4.67	6.7K	6.4K
DropIn	9.00	5.3K	7.2K
Membership	7.13	4.2K	702

Table 4: AL with human annotator results. ΔSER is % relative gain compared to the existing model. Higher is better.

Table 4 shows the results from AL with human annotators. On each feature, AL improved our existing NLU model by a statistically significant 4.6%-9%. On average 25% of utterances are false positive. This is lower than the 50% in the simulation because the initial training data exhibits more examples of the negative class. Around 10% of the AL selected data is lost due to being unactionable or out-of-domain, similar to the frequency with which these utterances are collected by random sampling.

While working with human annotators on new domains, we observed two challenges that impact the improvements from AL. First, annotators make more mistakes on AL selected utterances as they are more ambiguous. Second, new domains may have a limited amount of test data, so the impact of AL cannot be fully measured. Currently, we address the annotation mistakes with manual data clean up and transformations, but further research is needed to develop an automated solution. To improve the coverage of the test dataset for new domains we are exploring test data selection using stratified sampling.

5 Conclusions

In this work, we focused on AL methods designed to select live data for manual annotation. The difference with prior work on AL is that we specifically target new domains in NLU. Our proposed Majority-CRF algorithm leads to statistically significant performance gains over standard AL and random sampling methods while working with a limited annotation budget. In simulations, our Majority-CRF algorithm showed an improvement of 6.6%-9% SER relative gain compared to random sampling, as well as improvements over other AL algorithms with the same annotation budget. Similarly, results with live annotators show statistically significant improvements of 4.6%-9% compared to the existing NLU system.

References

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*.

Renato De Mori, Frédéric Bechet, Dilek Hakkani-Tur, Michael McTear, Giuseppe Riccardi, and Gokhan Tur. 2008. Spoken language understanding.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine learning*.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. 2013. *Nonparametric statistical methods*. John Wiley & Sons.

Hong-Kwang Jeff Kuo and Vaibhava Goel. 2005. Active learning with minimum expected error for spoken language understanding. In *INTERSPEECH*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*.

John Langford, Lihong Li, and Alex Strehl. 2007. Vowpal Wabbit.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the eleventh international conference on machine learning*.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *In Proceedings of DARPA Broadcast News Workshop*.

Naoki Abe Hiroshi Mamitsuka et al. 1998. Query learning strategies using boosting and bagging. In *Machine learning: proceedings of the fifteenth international conference (ICML'98)*, volume 1. Morgan Kaufmann Pub.

Thomas Osugi, Deng Kim, and Stephen Scott. 2005. Balancing exploration and exploitation: A new algorithm for active machine learning. In *Data Mining, Fifth IEEE International Conference on*.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*.

Hinrich Schütze, Emre Velipasaoglu, and Jan O Pedersen. 2006. Performance thresholding in practical text classification.

Burr Settles. 2009. Active learning literature survey. Computer sciences technical report, University of Wisconsin–Madison.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-instance active learning. In *Advances in neural information processing systems*.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*.

Chengwei Su, Rahul Gupta, Shankar Ananthakrishnan, and Spyros Matsoukas. 2018. A re-ranker scheme for integrating large scale nlu models. *arXiv preprint arXiv:1809.09605*.

Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*.

Gokhan Tur, Robert E Schapire, and Dilek Hakkani-Tur. 2003. Active learning for spoken language understanding. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*.