Deep Active Learning for Sequence Labeling Based on Diversity and Uncertainty in Gradient

Yekyung Kim
LG Electronics AI Lab
`yekyung.kim@lge.com`

Abstract

Recently, several studies have investigated active learning (AL) for natural language processing tasks to alleviate data dependency. However, for query selection, most of these studies mainly rely on uncertainty-based sampling, which generally does not exploit the structural information of the unlabeled data. This leads to a sampling bias in the batch active learning setting, which selects several samples at once. In this work, we demonstrate that the amount of labeled training data can be reduced using active learning when it incorporates both uncertainty and diversity in the sequence labeling task. We examined the effects of our sequence-based approach by selecting weighted diverse in the gradient embedding approach across multiple tasks, datasets, models, and consistently outperform classic uncertainty-based sampling and diversity-based sampling.

1 Introduction

Sequence labeling is one of the commonly used techniques for solving natural language understanding (NLP) tasks such as named-entity recognition (NER) and slot filling. Furthermore, for these tasks, the state-of-the-art results are typically based on deep neural networks (Kurata et al., 2016; Liu and Lane, 2016; Ma and Hovy, 2016). However, the performance of these models is highly dependent on the availability of large amounts of annotated data. Moreover, compared with classification tasks, which require only one label for a sample, the sequence learning tasks require a series of token-level labels for an entire sequence, which makes them time-consuming and a costly annotation process.

This problem can be mitigated using active learning (AL), which achieves improved performance with fewer annotations by strategically selecting the examples to annotate (Cohn et al., 1996; Settles and Craven, 2008; Siddhant and Lipton, 2018; Shen et al., 2018). There are two major strategies for active learning, namely, diversity-based sampling and uncertainty-based sampling (Muthakana, 2019; Ash et al., 2020). Traditionally, uncertainty-based sampling is the most common pool-based AL approach. However, previous work pointed out that focusing only on the uncertainty leads to a sampling bias (Dasgupta, 2011). It creates a pathological scenario where selected samples are highly similar to each other, which clearly indicates inefficiency. This may cause problems, especially in the case of noisy and redundant real-world datasets. Another approach is diversity-based sampling, wherein the model selects a diverse set such that it represent the input space without adding considerable redundancy (Sener and Savarese, 2018). This approach can select samples while ensuring a maximum batch diversity. However, this approach might select points that provide little new information, thereby reducing the uncertainty of the model. Certain recent studies for classification tasks implemented an algorithm named Batch Active learning by Diverse Gradient Embeddings (BADGE). This algorithm first computes embedding for each unlabeled sample based on induced gradients, and then geometrically picks the instances from the space to ensure their diversity (Ash et al., 2020). Although it proves to be a robust improvement when performing an image classification task, its performance in sequence labeling tasks is yet unproven. In this study, we investigated some practical active learning algorithms that consider uncertainty and diversity in sequence labeling tasks over different datasets and models. Moreover, we suggested a method to expand BADGE with weighted sampling based on the sequence length to ensure

Utterance	show	me	flight	from	Boston	to	Denver
Slot label	O	O	O	O	B-fromloc	O	B-toloc
Entity label	O	O	O	O	B-Location	O	B-Location

Table 1: An example of utterance and the annotation

cost-effective labeling. This simple modification in it has a positive implication that it tends to select cost-effective samples. The proposed model trades off between uncertainty and diversity by selecting diverse samples in the gradient space depending on the parameters in the final layer, for which, the currently available models focus only on uncertainty. To the best of our knowledge, our study is the first to apply diverse gradient embedding to a sequence labeling task. We experimented with the CoNLL 2003 English, ATIS, and Facebook Multilingual Task Oriented Dataset (FMTOD). Accordingly, it was empirically demonstrated that the proposed method consistently outperformed the baseline method including Bayesian AL by disagreement (BALD), which shows state-of-the-art performance in NER task, across the datasets, tasks and model architectures.

2 Related Work

Several recent papers investigated AL to alleviate the data dependency of deep learning for NLP. A different query criterion based on expected gradient length (EGL) has been proposed (Settles and Craven, 2008). Furthermore, Zhang et al. (2017) who addressed text classification, proposed selecting the examples according to the expected gradient length of the word embedding layer. Siddhant and Lipton (2018); Shen et al. (2018) addressed Deep Bayesian Active Learning for NER task and proved that BALD exhibits a state-of-the-arts performance. However, those works do not consider the diversity of the examples.

Shen et al. (2018) attempted solving this problem through a hybrid AL method that performed a representativeness-based sampling weighted by uncertainty. However, it could not outperform the uncertainty-based methods. For the slot filling task, researchers proposed adversarial AL for sequence learning with an additional discriminator network (Deng et al., 2018) and submodularity-inspired data ranking function (Dimovski et al., 2018) to select low-data regime. Some studies have explored hybrid methods that incorporate both diversity and uncertainty in classifica-

tion. Zhdanov (2019) takes into account both the informativeness of the examples for the model as well as their diversity in a mini-batch during classification; moreover, it considers uncertainty as a scalar value similar to works discussed earlier herein. Ash et al. (2020) also designed a BADGE method to incorporate both predictive uncertainty and sample diversity. This method exhibits a robust performance in different environmental settings while performing an image classification task. In this work, we have provided an empirical evaluation of diverse gradient embeddings for NLP tasks using different models, and explored a different method to consider variable-length input.

3 Background

3.1 Sequence labeling

Slot-filling and NER are the fundamental tasks for building spoken language understanding systems. Both of these utilize sequence labeling as an approach and serve different purposes. The slot-filling task finds relevant information in a query and tags each token with the corresponding slot labels. Comparatively, NER is more generic and is likely to have fewer labels than slot-filling. A example of this alignment comparison is provided in Table 1. In sequence labeling, the training examples map an input sequence, x, to the corresponding label sequence, y. The input of this sequence is comprised of a sequence of words, denoted as $x = (w_1, w_2, ..., w_t)$. Moreover, the input and label sequences have identical lengths. Therefore, there is an explicit alignment at each time step of the sentence. We used two architectures for training: the bidirectional long short-term memory (BiLSTM) tagger and the BiLSTM encoder-decoder, as described in Figure 1.

3.2 AL for sequence

There are some existing active learning approaches that already show promising performance for sequence labeling. Among these approaches, we considered two uncertainty-based methods and one diversity-based method as

(a) BiLSTM tagger without neural decoder layer
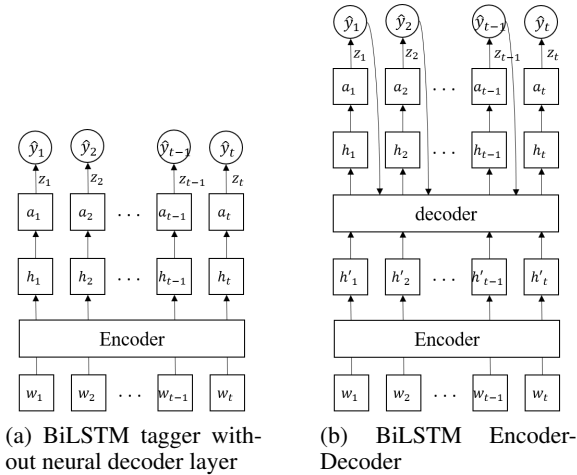
(b) BiLSTM Encoder-Decoder

Figure 1: recurrent neural network (RNN)-based architecture for sequence labeling

baseline. Specifically, the uncertainty based methods we reviewed are: Maximum normalized log-probability (MNLP) and BALD, which exhibit a state-of-the-art performance (Shen et al., 2018; Siddhant and Lipton, 2018). In addition, the diversity-based method we considered herein adopts a coreset algorithm, instead of performing a direct evaluation of the classification task (Sener and Savarese, 2018).

Maximum normalized log-probability (MNLP): MNLP extends the least confidence (LC) method used for sequence selection using log probability normalized by sequence length and removing the bias observed in the LC model, which preferentially selects longer sentences (Shen et al., 2018). Let n be the length of the sequence of x_i. Then,

$$\frac{1}{n} \sum_{i=1}^{n} \max_{y_i,..y_n} \log P(y_i|y_1,..,y_n,\{x_i\}) \quad (1)$$

Bayesian AL by disagreement (BALD): The Monte Carlo (MC) variant of BALD exploits an interpretation of dropout regularization as a Bayesian approximation to a Gaussian process (Gal et al., 2017; Siddhant and Lipton, 2018). During inference, a fixed number of forward passes is executed with a dropout. The measure of uncertainty is the fraction of models across MC dropout samples from the network that disagree with the most popular choice. Let $\widetilde{y}_j^1$ represent the prediction of tags applied to the tth forward pass on jth sample, and T be the number of forward passes.

Then,

$$\operatorname*{argmin}_{j} \left(1 - \frac{\operatorname{count}(\operatorname{mode}(\widetilde{y}_j^1, ..., \widetilde{y}_j^T))}{T} \right) \quad (2)$$

In this paper, we considered T = 100 independent dropout masks.

Core-set: Core-set is a pure diversity-based approach to find a subset of points, which is termed core-set, such that for all points, the maximum distance to the closest selected point is minimized using core-set selection (Sener and Savarese, 2018). While the core-set is proposed for the convolutional neural network in the classification task, we adopted it for sequence labeling tasks with an RNN network. We leveraged the greedy furthest-first traversal condition on all labeled examples within the embedding; moreover, each example was computed by the encoder layer.

4 Diverse Gradient Embedding for Sequences

In this section, we introduce a method that simultaneously captures both uncertainty and diversity for active learning with sequence tagging. BADGE only addresses the image classification task. Moreover, it does not consider tasks that have sequences. In this section, first, we discuss how the gradient embedding of the last layer is related to the representation information. Second, we propose a simple modified sampling algorithm with length-normalized weight for improved sampling efficiency.

Algorithm 1 Active learning with diverse gradient embeddings for sequence labeling tasks.

Input: Unlabeled dataset U, sequence length of example L, number of budgets in query K.

1. For all samples in U
    1. Compute the hypothesis label $\widehat{y}$
    2. Compute the gradient embedding g_x of the last layer
2. Select data using weighted k-means++ seeding algorithm with L on g_x until getting K number of samples.

Similar to Ash et al. (2020), our method, described in Algorithm 1, performs two main computations at each AL round: (i) extracting gradient embeddings with respect to the parameters of final

layers for all the unlabeled samples, and (ii) sampling a batch of query points based on these gradient embeddings using the weighted k-means++ initialization. It selects sentences for which the diversity and uncertainty are high when considering its length. The details of each computation to adapt it to the sequence labeling task is descried as follows:

4.1 Gradient Embedding

Here, we describe the gradient embedding of the penultimate layer in the sequence tagging task. Deep neural networks are optimized using gradient-based methods; therefore, training the gradient, back-propagated to a set of model parameters, captures the uncertainty of an example x, which is labeled with y. This may be viewed as a measure of change.

As the true label in the AL setting is unknown, Zhang et al. (2017) investigated AL for sentence classification with expected gradient length over all possible classes. We computed the gradient assuming that the model's current prediction on the example is its true label because its norm provides a lower bound on the gradient norm for the true label (Ash et al., 2020).

For given labels y and sample sequence x, this can minimize the negative log-likelihood of the objective function $L_{NL}(x, y)$ for the sequential labeling neural network. We denoted the nonlinear function that maps an input x to output h_t of the network's penultimate layer that has a hidden state of current observation at time t as z_t. In the last layer, $a_t = W_{hz}h_t + b_z$ where W_{hz} is the weights of the layer and b_z is bias. $z_t = softmax(a_t)$ is the nonlinearity softmax that predicts the probability assigned to the K classes.

$$L_{NL}(x, y) = \sum_t y_t log z_t \quad (3)$$

Note that the weight of the last layer, W_{hz} is shared across all time sequences; consequently, we can differentiate W_{hz} at each time step. If $g^i_{x_t}$ is defined as the derivative of the last layer weight, W_{hz}, for $L_{NL}(x, y)$, and the hypothesis label $\widehat{y}_t = argmax(z_t)$ is assumed, then the ith block gradient corresponding to label i is

$$(g_x)_i = \frac{\partial}{\partial W_{hz}} L_{NL}(x, \widehat{y}) = \sum_t -(\widehat{y}_{ti} - p_{ti})h_t \quad (4)$$

Based on Equation 4, each block of g_x is the sum of the product of the output of the penultimate layer and the probability vector at each time step. Although traditional representation learning in RNN uses only the final hidden state in the last step, the summation of each hidden state retains the property of representation information. In this respect, g_x captures sample x's representation and has the probability scaling. It measures uncertainty as the gradient magnitude with respect to the parameters in the final layer.

4.2 Diversity Based Sampling

To capture diversity, BADGE uses k-means++ initialization sampling (Arthur and Vassilvitskii, 2007) that favors both high magnitude and geometrical diversity based on gradient embeddings g_x. In particular, it performs sequential sampling of k centers, where each new center is sampled from the ground set with probability proportional to the squared distance $d(x)$ to its nearest center. Moreover, each sentence has the same weight for classification tasks because we measured the budget in sentences. In sequence labeling tasks, however, we considered that the cost of annotating is proportional to the number of words because the annotator must provide one tag per word and that every word in the selected sentence must be annotated at once (Shen et al., 2018). Therefore, it is important to select representative samples while considering the length of the sentence. Accordingly, we modified k-means++ initialization to weight the length of each sample.

Weighted K-means++: We modified the probability to sample the new center by weighted probability with the weight of the sentence length $w(x)$ of x data. Denoting the length of x data as $l(x)$, $w(x)$ can be expressed as $w(x) = l(x)^2 / \sum_x l(x)^2$. In other words,

$$p(x) = \frac{w(x)d(x)^2}{\sum_x w(x)d(x)^2}. \quad (5)$$

We abbreviated AL for diverse gradient embeddings with weighted k-means++ as W-BADGE.

5 Experimental Setting

5.1 Datasets

For the experiments, we used three publicly available datasets: CoNLL 2003 (Sang and Meulder, 2003) for the named entity recognition task, ATIS (Hemphill et al., 1990) and

Facebook Multilingual Task Oriented Dataset (Schuster et al., 2019) dataset for the slot-filling task. We followed the standard splits dataset train/validation/test. CoNLL 2003 English NER data contains 14,041/3,250/3,453 samples for the trainvalidation/test sets. This dataset contains four different types of named entities: PERSON, LOCATION, ORGANIZATION, and MISC. ATIS comprises conversations from the airline domains. It contains 4,478/500/893 utterances for train/validation/test sets with 11 types of slots. The FMTOD dataset comprises multiple domains including alarm, reminder, and weather. We used the English data in FMTOD, which have 30,521/4,181/8,621 datasets and 11 types of slots.

5.2 Model and Hyperparameter

The two architectures used for training were the BiLSTM tagger (BiLSTM for word-level encoding) and BiLSTM encoder-decoder (BiLSTM for word-level encoding, and LSTM for decoding) (Zhu and Yu, 2017) for both NER and slot-filling tasks, as shown in Figure 2. For the encoder, BiLSTM for word-level encoding contained 200 hidden units for ATIS and CoNLL, whereas it had 256 hidden units for FMTOD. Our model was comprised of 400-dimension word embeddings and utilized pre-trained GloVe and Kazuma (Hashimoto et al., 2017). Furthermore, we comprised a BiLSTM decoder with 100 dimensions for all datasets and utilized the greedy search for decoding. An Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and a dropout rate of 0.5 were used to train our entire deep learning system. Additionally, we used 20, 16, and 32 batches for ATIS, CoNLL, and FMTOD, respectively. We trained model for 50 epochs for ATIS and FMTOD, whereas this number was 25 for CoNLL.

5.3 Training Configuration

The AL process begins with initial samples randomly selected from the training dataset. We trained the initial model using this data. The learning process, which follows subsequently, consists of multiple rounds. At the beginning of each round, the AL algorithm selects sentences from the remaining training data to be annotated up to the predefined budget. After labeling the annotations, they are added to the training data. Accordingly, the model parameters were updated by

training it on the new training dataset before proceeding to the next iteration. In each round, we trained the model from scratch to prevent overfitting (Hu et al., 2019). We began our experiments with an initial labeled pool with 2% labels of original training data for the CoNLL and ATIS datasets, and 1% labels of those for the FMTOD. Further, we added the same number of labels at each iteration of active learning and evaluated the performance of the algorithm by its F1 score on the test dataset. We also reported the performance achieved after the full training of our model. All experiments were repeated five times and the average F-scores with their standard deviations are reported.

6 Results

We evaluated the performances of various AL methods with different models for the slot-filling and NER tasks and plotted them in Figures 2 and 3, respectively. The x-axis represents the percent of words annotated and used for training and the y-axis indicates the best F1 scores obtained. Moreover, standard errors are indicated by shaded regions. In all cases, we observed that the W-BADGE method indicated a significant improvement or over the baseline AL methods. It consistently outperforms those employing either pure uncertainty or diversity-based sampling in both the earlier and latter rounds. The performance gap is clearer, especially in the earlier rounds of training. Although BADGE is comparable with other methods, W-BADGE performs as well as or better than BADGE.

More specifically, as can be observed from Figure 2, W-BADGE performs satisfactorily in both the ATIS and FMTOD datasets and shows significant improvement in all baselines when performing the slot-filling task. While BADGE also shows the similar performance in FMTOD dataset with BiLSTM tagger, W-BADGE slightly outperformed in BiLSTM Encoder-Decoder model. Moreover, classic uncertainty-based sampling MNLP outperforms BALD and Coreset in both datasets. We suppose that the slot-filling task is relatively complex and sparse because it has more classes than the NER task; therefore, uncertainty-based sampling methods have an advantage when performing such a task.

Figure 3 graphs the learning curves of the AL algorithms on CoNLL 2003 dataset for NER task.
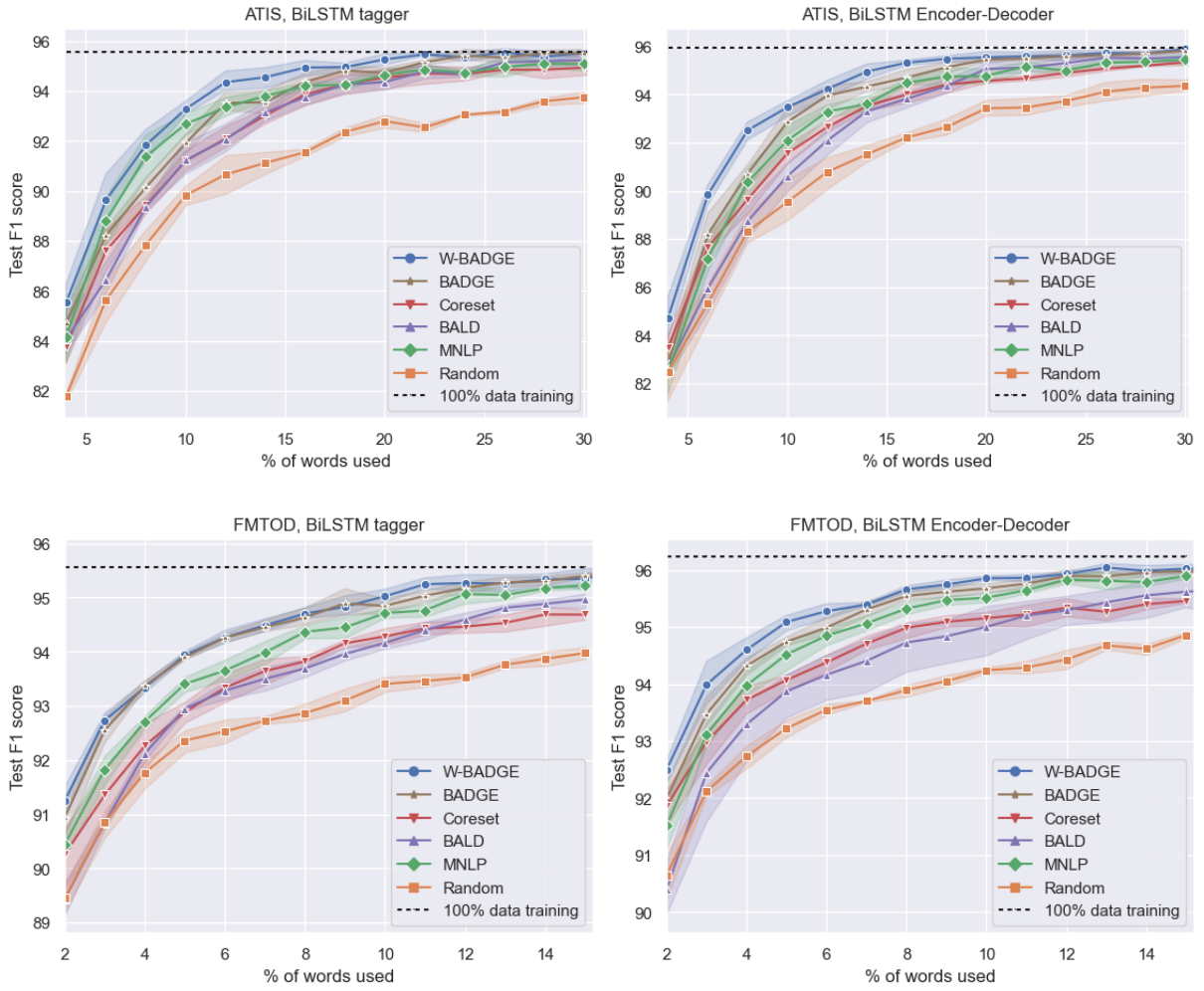
Figure 2: Performance comparison of different datasets and various AL sampling methods used for the slot-filling task
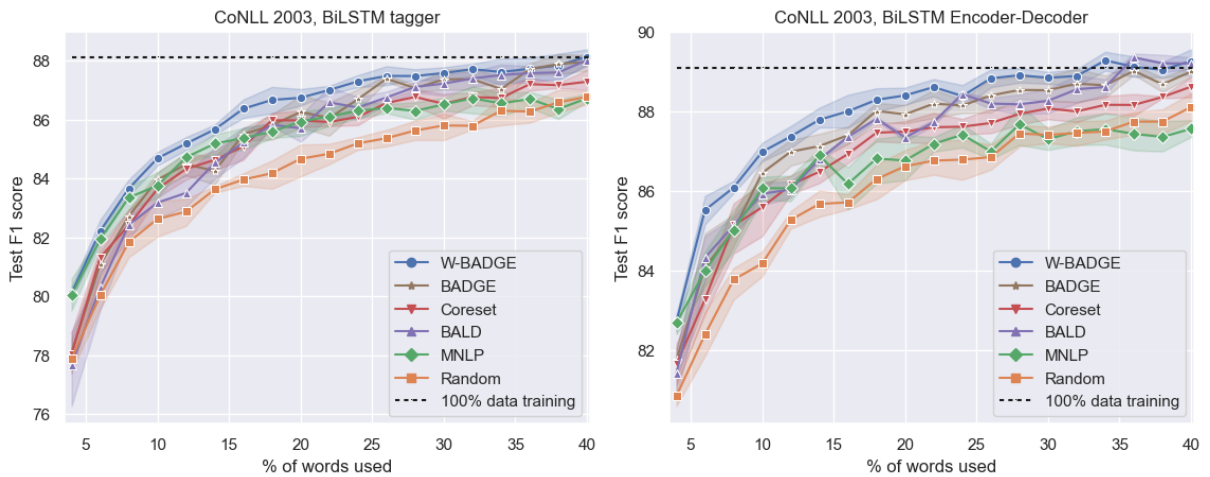

Figure 3: Performance comparison of CoNLL2003 dataset and various AL sampling methods in the case of NER task

We observed that W-BADGE presented significant   improvement over BALD, which shows the best

performance in NER tasks (Siddhant and Lipton, 2018) and BADGE. As W-BADGE outperformed the Coreset, it appears that W-BADGE can be advantageous not only over purely diversity-based approaches but also over classic uncertainty-based approaches in NER task.

Overall, the advantages of weighted diverse gradient embeddings can be substantial. For example, We find that active learning algorithms achieve 98-99% deep model trained on full data using only 15% of samples in ATIS dataset. The relative improvement remains significant over baselines.

7 Conclusion

In this study, we explored the empirical study on AL utilizing the advantages of both uncertainty and diversity by selecting weighted diverse gradient embeddings to perform a sequence labeling task. We proposed an efficient method and empirically demonstrated that it could consistently achieve a superior performance while consuming much less data. It adds robustness to the dataset and the architecture, thus proving to be a useful option for solving real-world active learning problems

References

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *J. Artif. Intell. Res.*, 4:129–145.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781.

Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. 2018. Adversarial active learning for sequences labeling and generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4012–4018. ijcai.org.

Mladen Dimovski, Vladimir Ilievski, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2018. Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings. *CoRR*, abs/1802.00757.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1923–1933. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, USA, June 24-27, 1990*. Morgan Kaufmann.

Peiyun Hu, Zachary C. Lipton, Anima Anandkumar, and Deva Ramanan. 2019. Active learning with partial feedback. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2077–2083. The Association for Computational Linguistics.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 685–689. ISCA.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Hariank Muthakana. 2019. Uncertainty and diversity in deep active image classification. Master's thesis, Carnegie Mellon University, 12.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3795–3805. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1070–1079. ACL.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2904–2909. Association for Computational Linguistics.

Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active discriminative text representation learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3386–3392. AAAI Press.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *CoRR*, abs/1901.05954.

Su Zhu and Kai Yu. 2017. Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5675–5679. IEEE.