Leveraging Active Learning to Minimise SRL Annotation Across Corpora

Skatje Myers
University of Colorado at Boulder
skatje.myers@colorado.edu

Martha Palmer
University of Colorado at Boulder
mpalmer@colorado.edu

Abstract

In this paper we investigate the application of active learning to semantic role labeling (SRL) using Bayesian Active Learning by Disagreement (BALD). Our new predicate-focused selection method quickly improves efficiency on three different specialised domain corpora. This is encouraging news for researchers wanting to port SRL to domain specific applications. Interestingly, with the large and diverse *OntoNotes* corpus, the sentence selection approach, that collects a larger number of predicates, taking more time to annotate, fares better than the predicate approach. In this paper, we analyze both the selections made by our two selections methods for the various domains and the differences between these corpora in detail.

1 Introduction

The majority of natural language processing (NLP) systems are reliant on manual annotations to train supervised models. Although semi-supervised and unsupervised methods are frequently employed to help adapt models to new domains, human annotation remains the gold standard for quality input. Due to the high cost of human annotation, especially if the task requires expert knowledge, and the time-intensive process, this can be daunting for many applications.

Active learning (AL) has been shown to reduce annotation requirements for a variety of NLP tasks (Zhang et al., 2022) by selecting more informative instances that are most likely to fill gaps in the model's knowledge.

In this paper, we focus specifically on the NLP task of semantic role labeling (SRL). The goal of SRL is to identify and label the who, what, and when of predicates in a sentence. This information can be used as features in downstream applications such as information extraction (MacAvaney et al., 2017), machine translation (Marcheggiani et al., 2018), and features prominently in Abstract Meaning Representation (AMR) applications (Banarescu et al., 2013).

In this paper, we propose a new selection strategy tuned for SRL that is based off of previous methods of using model dropout to approximate a Gaussian process (Siddhant and Lipton, 2018). We compare this to prior work on AL selection for SRL (Myers and Palmer, 2021) on four corpora in a variety of domains: ecology, earthquakes, clinical notes, and the large multi-genre *OntoNotes* corpus.

Since sentences in most domains typically contain multiple predicates, there are often redundancies in choosing predicates to annotate on the sentence level. Although a sentence may contain a particularly informative predicate, annotating high-frequency verbs such as "be" that co-occur in the sentence may not be beneficial. We instead use a method to select specific predicate-argument structures and compare the impact on performance as compared to selecting whole sentences instead.

This method is a natural extension that allows us to even better leverage the focused annotation that active learning offers by using a more granular approach. While we find consistent early benefit in the more domain-specific corpora, this finer-grained approach proves to be slower for the more diverse *OntoNotes*.

We also explore the statistical differences between these corpora, the selections our algorithm makes, and test a variety of selection batch sizes in order to shed light on expectations for use in future domains.

2 Background

Proposition Bank (PropBank) (Palmer et al., 2005) is verb-oriented semantic representation consisting of a predicate and its arguments. Predicates are given a *roleset ID*, which distinguishes the sense of the word, such as play.01 (*to play a game*) or play.02 (*to play a role*). Each roleset has its own list of permissible semantic roles, or arguments, for

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 399–408
July 13-14, 2023 ©2023 Association for Computational Linguistics

play.01	
play a game	
ARG0	player
ARG1	game
ARG2	equipment
ARG3	opponent

Table 1: PropBank roleset for *play.01*

that predicate, such as ARG0 (typically the agent of the action). Additionally, all rolesets support the use of a set of modifier arguments such as location (ARGM-LOC) and direction (ARGM-DIR). These arguments are annotated for the constituent spans of the sentence. For example:

[$_{\text{ARG0}}$ I] [$_{\text{Pred}}$ played] [$_{\text{ARG1}}$ chess] [$_{\text{ARG3}}$ against him].

Active learning is an iterative process by which data is selected for annotation using the model's own confidence. After initially training the model on a small amount of annotated data (referred to as the seed set), each unlabeled instance is predicted by the model and those that the model is least certain about (conventionally, by the model's outputs) are presumed to be more informative to learn from than those that the model has high certainty about. The uncertain instances can then be manually annotated and added into the training pool for the next training iteration. This process can repeast until either the performance is no longer significantly increasing or time/budget has been exhausted.

Previous work has shown that neural networks tend to be overconfident in their predictions, owing to their nonlinearity and tendency to overfit (Gal and Ghahramani, 2016)(Dong et al., 2018). Therefore, more recent work (Siddhant and Lipton, 2018) (Shen et al., 2017) has explored using Bayesian Active Learning by Disagreement (Houlsby et al., 2011) (BALD) rather than model outputs as a way of selecting informative instances for active learning for SRL and other NLP tasks. By using dropout during prediction, multiple forward passes can be treated as Monte Carlo draws from a stochastic model. The instances that have more disagreement amongst the predictions are considered to be more informative for the model to learn from.

Myers and Palmer (2021) applied BALD to SRL by calculating disagreement among five forward passes of the trained model using dropout, breaking down agreement scores by individual argument labels. We describe this in more detail in Section 4.1. The active learner used two alternative methods to select sentences: 1) using the average disagreement score amongst all predicates in the sentence (BALD-AP) or 2) by choosing the sentences that contain the single lowest scoring predicate (BALD-LSP). Since BALD-LSP performed best, we compare our predicate-focused BALD strategy against this method on both corpora used previously (*OntoNotes* and *THYME Colon*) as well as two new geoscience corpora from the ClearEarth project (Duerr et al., 2016).

3 Data

We aim to provide a demonstration of active learning for SRL across a variety of domains and sublanguages (Kittredge, 1982). Some knowledge domains exhibit narrow lexical, syntactic, and semantic structures that distinguish them from more general-purpose domains. This can lower performance dramatically when testing with an off-the-shelf general purpose model. Special techniques that take these domain specific-structures into account are needed for adapting NLP tools to these domains, as illustrated below.

THYME Colon is comprised of unstructured clinical notes relating to treatment of colon cancer (Albright et al., 2013). This corpus contains specialised medical vocabulary for a narrow domain and a large number of formulaic sentences, such as the following example:

> Pathology demonstrated a tubular adenoma with moderate dysplasia.

This contains medical terminology (tubular adenoma, dysplasia) as well as a non-standard use of *demonstrate*, which includes the shortening of *The pathology report* to simply *pathology*. This particular framing re-occurs frequently in THYME Colon, sometimes with *show* or *reveal* instead, and occasionally including the word *report* as in *pathology report*.

We also used two distinct geoscience domains from the ClearEarth project (Duerr et al., 2016):

- *Earthquakes* consists of 41k tokens of text from Wikipedia and education texts, and a glossary. This text includes specialised scientific language relating to earthquakes and plate tectonics, but also discussion of the history of the field at a high school reading level and content related to disasters. For example: *The ways that plates interact depend on their relative motion and whether oceanic or continental crust is at the edge of the lithospheric plate.*

- *Ecology* consists of 83k tokens of text from Wikipedia, educational websites, an ecology glossary, and Encyclopedia of Life. The scientific content covers genetics, evolution, reproduction, and food chains. For examples: *Anguis fragilis is an example of ovo-viviparity.* and *Alternatively, transcription factors can bind enzymes that modify the histones at the promoter.*

OntoNotes 5.0 (Weischedel et al., 2013) spans multiple genres, largely consisting of news sources, but also including telephone conversations, text from the New Testament, weblogs, and Usenet. This popular corpus serves as a broad purpose corpus for us, as opposed to the other more specialised domains.

We use a version of *OntoNotes* that does not include files that had no manual PropBank annotation performed. There still exist sentences within this version of the data that had only partial annotation, but we consider this to have a relatively small impact on performance.

Evaluation was performed on the standard test subset for each respective corpus.

4 Methods

We simulated active learning using AllenNLP's (Gardner et al., 2018) implementation of a state-of-the-art BERT-based SRL model (Shi and Lin, 2019).

In order to simulate active learning on each of these corpora, we partitioned the training subset of each corpus into 200 random sentences for seeding the learner, with the remainder used as the initial "unlabeled" pool for selection. The initial 200 seed sentences were the same across the three selection methods tested for each respective corpus.

After initially training on the seed set, we then select a batch of either 100 predicates or a number of sentences containing approximately 100 predicates to add to the training pool using the BALD PREDICATES or BALD SENTENCES strategy described below in Section 4.1 or by choosing 100 random predicates to simulate a passive learning approach. Results are reported on the test subset of the respective corpora and the model was retrained with the extended training pool. We continue iterations of selection and re-training until either all the data has been selected and moved into the training pool, or the experiment performances have sufficiently plateaued.

Our training procedure for this model used 25 epochs or stopped early with a patience of 5 based on the validation data for the relevant corpus.

4.1 Selection Methods

We use the BALD-LSP method tuned for SRL as described in Myers and Palmer (2021), which we will refer to in this paper as BALD SENTENCES for comparison.

After a model is trained, this method uses 10% dropout during 5 forward passes in order to generate multiple predictions for each instance in the unlabeled pool. For each predicate-argument structure in a sentence and each argument label type present in the predictions, we calculate how many of the 5 predictions do not match the mode predicted span. If all five predictions have different spans for an ARG1, for example, then this results in the highest possible disagreement score for ARG1.

After disagreement scores are calculated for each argument label, these scores are averaged to produce a score for the predicate. If there is only one predicate in a sentence, this is the score for the sentence. If a sentence has multiple predicates, the sentence is assigned the score of the predicate that had the most disagreement. The sentences with the highest scores are selected to be included in the next round of training.

Our BALD PREDICATES method is a more granular extension of this previous work. We use the same idea of scoring individual argument spans based on agreement and averaging them into a single score for a given predicate instance, but we do not combine the scores of all predicates within a given sentence. We instead use the score to choose specific predicate instances to add to the training pool

We also compare these two active learning methods against a passive baseline of selecting random predicate instances.

5 Results

We present the learning curves of the different selection methods for the four corpora are presented

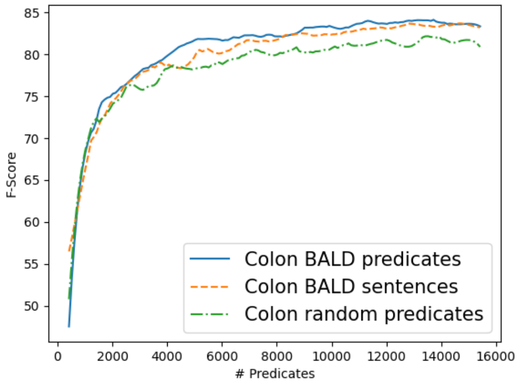

Figure 1: Performance of selection method by approximate number of predicates in the training pool on *THYME Colon* dataset.

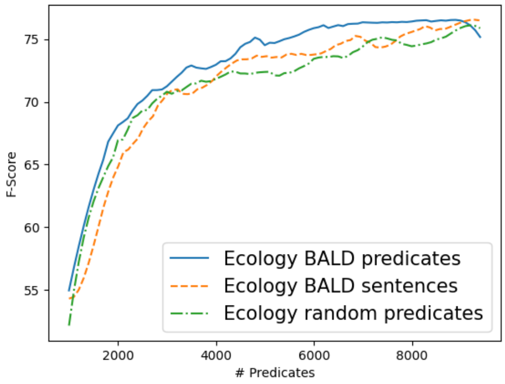

Figure 2: Performance of selection method by approximate number of predicates in the training pool on *ClearEarth Ecology* dataset.

in Figures 1, 2, 3, and 4. Natural variability in training the model produces some amount of noise, most prominently during the early iterations. In order to improve readability of these learning curves, we applied a Savitzky–Golay filter using a window of 15 data points and using a cubic polynomial.

We see consistent benefits of the BALD PREDICATES method at different points depending on the corpus.

For *Colon*, *Ecology*, and *Earthquakes* we begin to see consistent improvement for the BALD PREDICATES method over the other methods by approximately 1,500-2,000 predicates. On the other hand, for *OntoNotes*, it only catches up to random selection around 4,500 predicates and begins to improve over it around 7,000 predicates. For this corpus, BALD SENTENCES performs better.

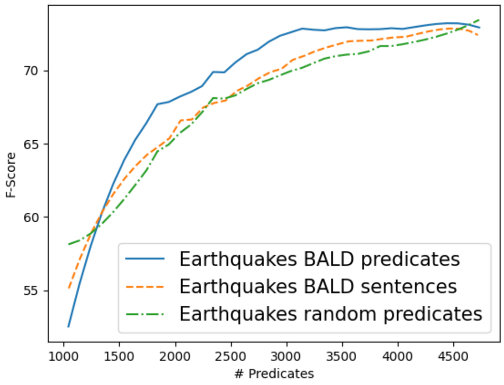

Figure 3: Performance of selection method by approximate number of predicates in the training pool on *ClearEarth Earthquakes* dataset.

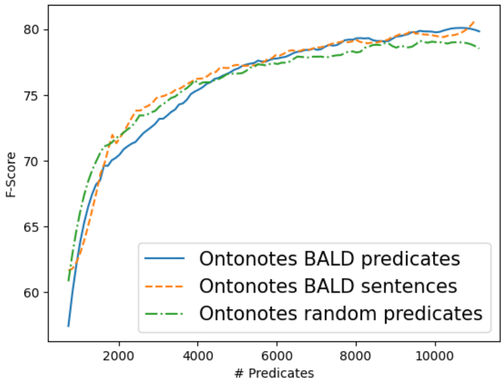

Figure 4: Performance of selection method by approximate number of predicates in the training pool on *OntoNotes*.

6 Analysis of Selections

In order to better understand the differences between the selection processes used and their variance across datasets, we examine the selections within each batch.

6.1 Diversity

By selecting multiple predicates or sentences in each iteration, we expect that there may be redundancies. For example, if the model has never seen a given predicate, it will likely have low confidence in its predictions for it. We present a study of the diversity of the selections over time.

We first observe the amount of redundancy within BALD PREDICATES. This method is choosing multiple instances of the same predicate lemma, as observed in Figure 5. In the two *ClearEarth* corpora we have analysed in this regard, which both ran to completion on the training data, approximately 25 of the 100 predicates are duplicates in the early phase of active learning and with redundancy getting worse as the process gets closer to completion. The results for *Colon* contain approximately similar amounts of redundancy for the duration we trained it.

While there may sometimes be value in selecting the same lemma in order to obtain multiple senses of the same predicate, minimising this could prove beneficial. Future work could be done to study the effect of limiting the selection batch to unique lemmas.

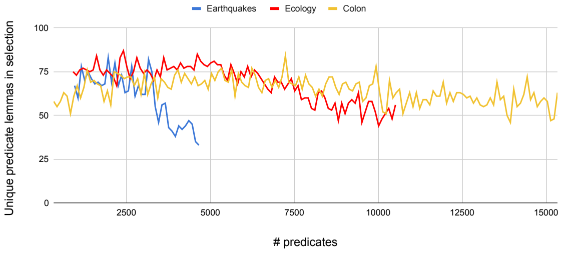

Figure 5: Number of unique predicate lemmas selected in each batch by the BALD PREDICATES method over iterations.

Additionally, the BALD PREDICATES method is capable of selecting multiple instances from the same sentence. While this may be beneficial, it's also possible that learning from just one predicate in the sentence will provide information that can improve agreement on other instances in the sentence.

We have found that for *Colon*, a randomly selected batch of 100 predicates contains 3 duplicate sentences on average, while the selections by BALD PREDICATES contain only 1 duplicate on average. For the *Ecology* corpus, both methods pick 3 duplicate sentences on average. This appears indicative that this is not a significant factor that necessitates correction.

Furthermore, we are interested in the sentence-level semantic redundancies within batches. Using the pre-trained all-mpnet-base-v2 model (Song et al., 2020), we can calculate the average pairwise cosine similarity between the unique sentences within batches. In Figure 6, we find that both active learning methods contain more sentence-level similarity on average (0.26) than what is chosen through random selection (0.19) from the *THYME Colon* corpus.

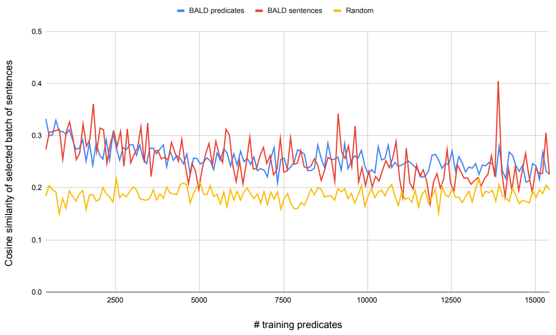

Figure 6: Average pairwise cosine similarity of selected sentences in each batch over iterations on *THYME Colon*.

We can see clear signs of the active learner choosing sentences that would be wasteful to have annotated. In one such batch, BALD SENTENCES selected 29 out of the 52 sentences where the sentences were all of the same basic form, but with varying AJCC cancer staging designations:

With available material: AJCC ypT1N0MX
With available surgical material [AJCC pT3N2Mx]

On the other hand, the difference in selection diversity is less pronounced on the other datasets. In Figure 7, we show the similarity in the selections on *ClearEarth Ecology*, where all methods average 0.20 across the iterations.

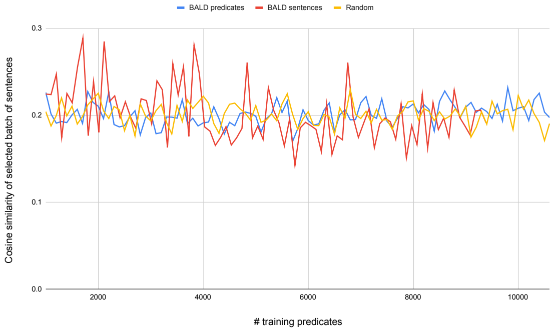

Figure 7: Average pairwise cosine similarity of selected sentences in each batch over iterations on *ClearEarth Ecology*.

6.2 Vocabulary Coverage

We hypothesised that a contributor to BALD PREDICATES's performance may be a rapid coverage of vocabulary, as predicates that involve unseen vocabulary could result in more disagreement. In Figure 8, we show the percentage of the unique vocabulary of the training set that is within the training pool as selections are made.

Across the datasets, we see varying results in how much BALD PREDICATES expedites vocabulary coverage. We find that BALD PREDICATES is not tending to choose unseen vocabulary compared to selecting predicates randomly for *Ecology*. On the other hand, active learning greatly accelerates this for *Ontonotes*, even after performance has largely plateaued. For *THYME Colon*, active learning provides an initial boost to vocabulary, but around the time that the performance plateaus, this decelerates below random.

6.3 Disagreement

For BALD PREDICATES, we calculate an average disagreement score for each selected batch. While early batches primarily contain predicates for which all predictions are in full disagreement, we see this disagreement trend downwards as performance plateaus. This is presented in Figure 9.

Although performance on *OntoNotes* has largely plateaued around an F-score of 79 by 7.5k training predicates, we know that training this model on the full dataset yields another 4 points. Since the disagreement scores of batches chosen by BALD PREDICATES is still over 70%, this seems indicative of the additional further performance to be gained, albeit at a slow pace that gets little value for the effort. In contrast, *Colon* plateaued around 82, but the benefits of annotating the remaining 50k predicates only provides an additional increase of 1 point. With the disagreement score having fallen below 45%, this points toward an appropriate stopping point.

7 Corpus Analysis

Although the new predicate selection method offers immediate benefit over BALD SENTENCES for the three sublanguage corpora, this is inconsistent with the result on *OntoNotes*, where selecting BALD SENTENCES is more advantageous until about 7k predicates. In order to better understand the possible reasons for this, we compare the make-up and distribution of the corpora. These statistics are presented in Table 2.

We use PropBank roleset ID's as our measure of polysemy, since we have gold standard annotation for them in all 4 corpora. Note that PropBank sense distinctions are fairly coarse-grained and were generally only created when there were differences between senses with respect to the semantic roles. VerbNet (Schuler, 2005), FrameNet (Baker et al., 1998) and WordNet (Miller, 1995) would all give much higher polysemy counts.

The largest and most diverse corpus in our experiments is *OntoNotes*, although we find that in terms of ratio of total tokens to predicates, unique rolesets, and unique tokens, *OntoNotes* is statistically more similar to the *THYME Colon Cancer* corpus than to either of the *ClearEarth* corpora. *OntoNotes* and *Colon* contain approximately one unique roleset per 376-403 tokens, whereas *Earthquakes* and *Ecology* contain one per 39 and 60 tokens, respectively.

Since *OntoNotes* covers a wider diversity of text types, it's unsurprising that it contains a much more diverse set of senses compared to the other corpora. While a lemma like "take" shows up with 25 different senses in *OntoNotes*, it only shows up in 8 senses in *Colon*.

For *OntoNotes*, only 30% of predicate occurrences are monosemous within the context of the corpus, whereas this figure is between 54%-61% for the other three corpora. 6% of the unique predicate lemmas within *OntoNotes* are seen in 3 or more rolesets, while this is true of only 2% of the set of lemmas in each of the other corpora.

We believe this polysemy factor may contribute to the predicate selection method being disproportionately slower to improve the learning curve on *OntoNotes* compared to the more focused domain corpora. BALD PREDICATES may be disadvantaged by more frequently choosing these rare senses even though they make up proportionally less of the training data and provide less value in terms of performance, but further investigation is needed.

	OntoNotes	Colon	Earthquakes	Ecology
Tokens	2.2 mil	522k	41k	83k
Unique tokens per token	44.55	36.88	8.42	10.43
Predicates	301k	57k	7.5k	15k
Tokens per predicate	7.41	9.11	39.63	60.45
Avg sentence length	18.74	11.33	23.39	24.48
Unique rolesets	5535	1389	1046	1376
Tokens per roleset	403	376	39	60
Predicate lemmas with 1 roleset	3829 (83.33%)	1340 (90.24%)	985 (91.20%)	1416 (92.73%)
Predicate lemmas with 2 rolesets	494 (10.75%)	112 (7.54%)	73 (6.76%)	80 (5.24%)
Predicate lemmas with 3+ rolesets	272 (5.92%)	33 (2.22%)	22 (2.04%)	31 (2.03%)
Monosemous predicate occurences	29.95%	55.02%	53.53%	60.94%

Table 2: Statistics about the four corpora.

8 Batch Sizes

Each iteration of active learning includes selecting an arbitrary number of instances to query. The number may be static, or dynamic with larger batches

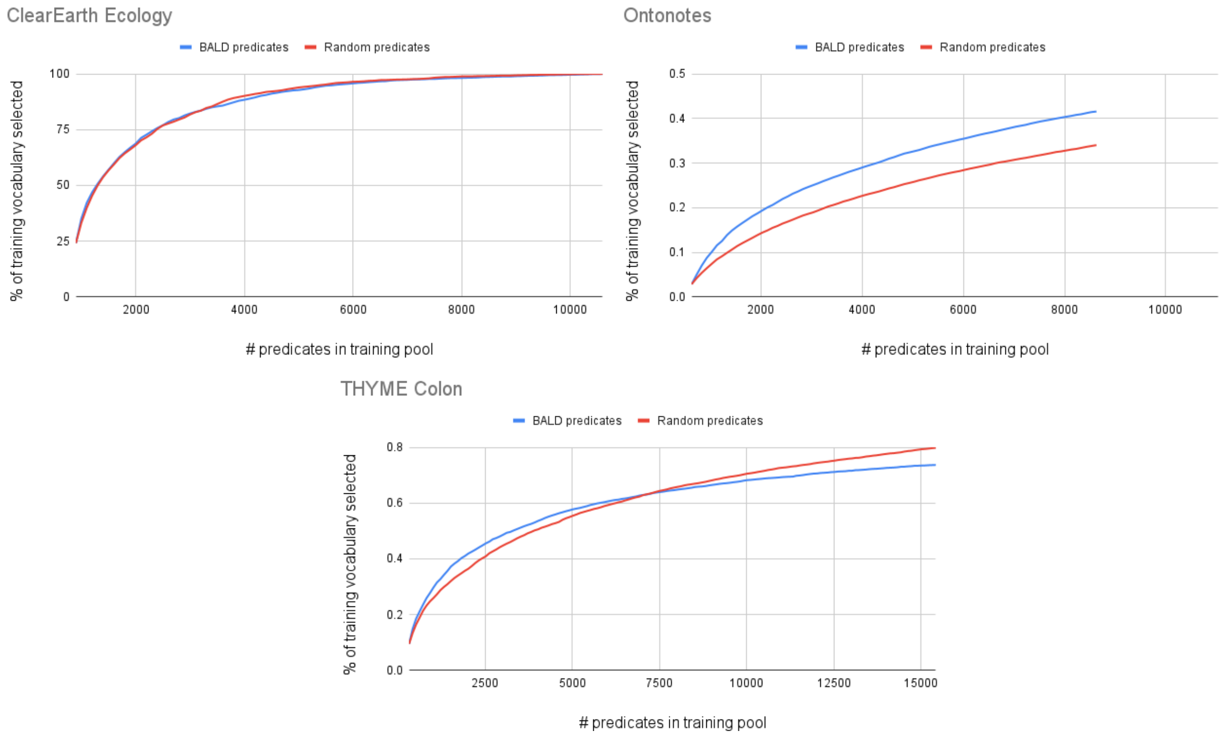

Figure 8: Percent coverage of training vocabulary in by number of predicates in training pool.

being selected in the early training process and smaller batches later on.

To maximally benefit from the model's feedback, in an ideal setup, each iteration would query for only one new instance, thereby minimizing the likelihood of selecting a batch of sentences with redundant information (Schohn and Cohn, 2000). Unfortunately, this leads to the process of active learning being significantly slower due to needing to re-train a model more often. Additionally, annotating a sentence at a time with long breaks in between may cost additional time on the part of the annotator due to mental context-switching and needing to load up appropriate software and resources. It would be more efficient for them to be able to annotate numerous examples in a row.

Our previous experiments testing the BALD PREDICATES method show positive results when selecting 100 predicates in a batch. This small batch size requires about 60 iterations before the learning curve plateaus for the Colon corpus. We examine the effect of larger batches on the learning curves for the *THYME Colon* and the two *ClearEarth* corpora.

8.1 Results

We used the BALD PREDICATES selection strategy with varying sizes of 100, 500, and 1000 query instances. These results are presented for three datasets in Figure 10, using datapoints on intervals of 1000 predicates.

Interestingly, changing the batch size has differing impacts on the datasets we examined this for. The *THYME Colon* corpus suffers very little from scaling all the way to 1000 predicates per selection batch. The results on Earthquakes show the clearest need for small batch sizes, while Ecology exhibits shifting performance over the course of iterations.

9 Conclusion and Future Work

In this paper, we've demonstrated that active learning can reduce annotation requirements for semantic role labeling across multiple domains by employing Bayesian Active Learning by Disagreement and using dropout to provide variability in predictions from the model. These predictions can be used to estimate the model's confidence in its predictions and select informative training instances to annotate.

Selecting predicate instances through the BALD PREDICATES method offers significant improvement in efficiency for *THYME Colon*, *ClearEarth Earthquakes* and *Ecology*, which have very focused domains. This method does not provide the same performance increase on the more general *OntoNotes* over the previous BALD SENTENCES, which selects whole sentences.

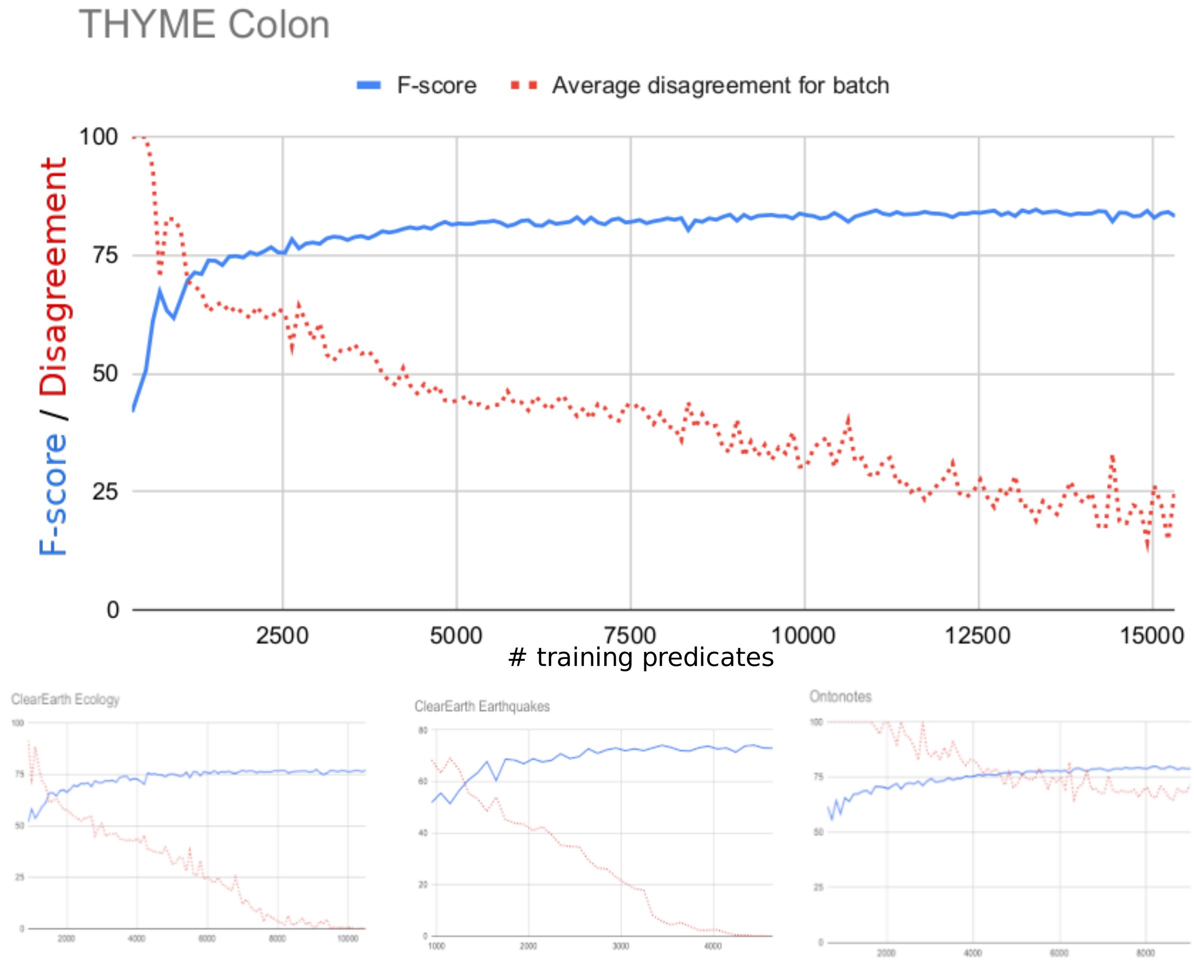

Figure 9: Average disagreement in selected batches decreases as iterations continue, while F-score increases and plateaus.

We have provided a statistical comparison of these corpora and offered some possible reasons for the divergence in performance, including a notable difference in polysemy within *OntoNotes* compared to the rest of the corpora.

Additionally, we examined the diversity of the selected predicates and sentences for BALD PREDICATES. Although these results vary across the different datasets, it indicates a couple potential avenues of future improvement. Reducing sentence-level semantic similarity seems of particular relevance to the *THYME Colon* corpus. We have also identified redundancies in the predicates chosen in each batch by BALD PREDICATES.

We also presented the change in model prediction disagreements over iterations as compared to model performance, which could be beneficial to determine when the costs of further annotation outweigh the additional gains that the model can provide.

Since the choice of how many selections to take on each iteration cannot be tuned for in real-world use of active learning, we have attempted to shed light on the levels of impact to expect on several different corpora, which vary in how sensitive they are to larger batches. We find that further investigation is needed to determine the most significant factors causing these differences so that future applications of active learning to SRL can predict the most ideal selection batch size that balances performance against training time for their target domain.

Limitations

While it reduces annotation costs, AL can be computationally intensive and its success is correlated to the number of training iterations. Whether this will be a net savings for a given project may vary from case to case, depending on computing resource availability and annotator costs. The work-

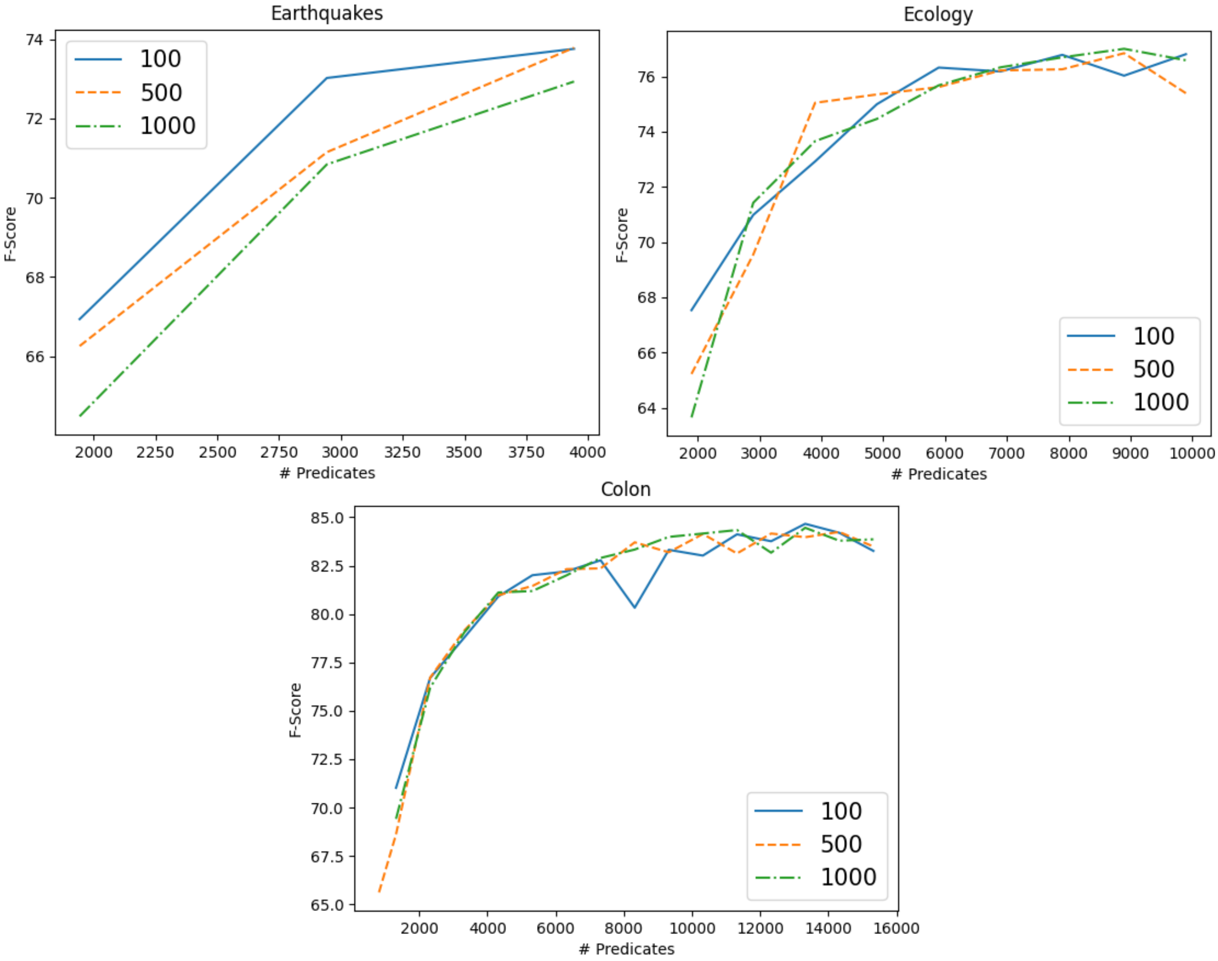

Figure 10: Performance of using BALD PREDICATES, selecting varying numbers of predicates per iteration.

flow of annotating and re-training may not be feasible in the budgetary constraints that inherently make AL desirable over randomly annotating training data.

Partial SRL annotation of sentences or documents may not be desirable in projects that simultaneously annotate other things, such as AMRs or coreference, which rely on whole-sentence or whole-document annotation.

Acknowledgments

We gratefully acknowledge the support of a sub-award from RPI on DARPA KAIROS Program No. FA8750-19-2-1004, a sub-award from Harvard Children's on NIH: 5R01LM010090-09 (THYME-3, Temporal Relation Discovery for Clinical Text), and NSF ACI 1443085: DIBBS (Porting Practical NLP and ML Semantics from Biomedicine to the Earth, Ice and Life Sciences). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, NSF, NIST, or the U.S. government.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, IV Styler, William F, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

R. Duerr, A. Thessen, C. J. Jenkins, M. Palmer, S. Myers, and S. Ramdeen. 2016. The ClearEarth Project: Preliminary Findings from Experiments in Applying the CLEARTK NLP Pipeline and Annotation Tools Developed for Biomedicine to the Earth Sciences. In *AGU Fall Meeting Abstracts*, volume 2016, pages IN11B–1625.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Richard Kittredge. 1982. Sublanguages. *American Journal of Computational Linguistics*, 8(2):79–84.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Skatje Myers and Martha Palmer. 2021. Tuning deep active learning for semantic role labeling. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 212–221, Groningen, The Netherlands (online). Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 839–846, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

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*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.