A Survey of Active Learning for Natural Language Processing

Zhisong Zhang, Emma Strubell, Eduard Hovy
Language Technologies Institute, Carnegie Mellon University
zhisongz@cs.cmu.edu, strubell@cmu.edu, hovy@cmu.edu

Abstract

In this work, we provide a literature review of active learning (AL) for its applications in natural language processing (NLP). In addition to a fine-grained categorization of query strategies, we also investigate several other important aspects of applying AL to NLP problems. These include AL for structured prediction tasks, annotation cost, model learning (especially with deep neural models), and starting and stopping AL. Finally, we conclude with a discussion of related topics and future directions.

1 Introduction

The majority of modern natural language processing (NLP) systems are based on data-driven machine learning models. The success of these models depends on the quality and quantity of the available target training data. While these models can obtain impressive performance if given enough supervision, it is usually expensive to collect large amounts of annotations, especially considering that the labeling process can be laborious and challenging for NLP tasks (§3.2). *Active learning* (AL), an approach that aims to achieve high accuracy with fewer training labels by allowing a model to choose the data to be annotated and used for learning, is a widely-studied approach to tackle this labeling bottleneck (Settles, 2009).

Active learning has been studied for more than twenty years (Lewis and Gale, 1994; Lewis and Catlett, 1994; Cohn et al., 1994, 1996) and there have been several literature surveys on this topic (Settles, 2009; Olsson, 2009; Fu et al., 2013; Aggarwal et al., 2014; Hino, 2020; Schröder and Niekler, 2020; Ren et al., 2021; Zhan et al., 2022). Nevertheless, there is still a lack of an AL survey for NLP that includes recent advances. Settles (2009) and Olsson (2009) provide great surveys covering AL for NLP, but these surveys are now more than a decade old. In the meantime, the field of NLP has been transformed by deep learning. While

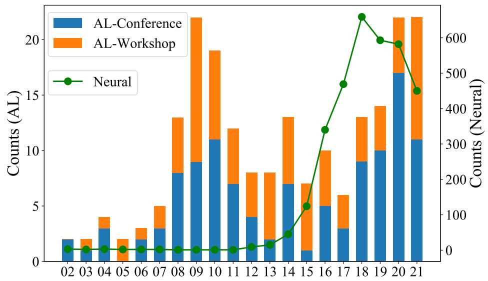

Figure 1: Counts of AL (left) and *"neural"* (right) papers in the ACL Anthology over the past twenty years.

other more recent surveys cover deep active learning, they are either too specific, focused only on text classification (Schröder and Niekler, 2020), or too general, covering AI applications more broadly (Ren et al., 2021; Zhan et al., 2022). Moreover, applying AL to NLP tasks requires specific considerations, e.g. handling complex output structures and trade-offs in text annotation cost (§3), which have not been thoroughly discussed.

In order to provide an NLP-specific AL survey,[1] we start by searching the ACL Anthology for AL-related papers. We simply search for the keyword *"active"* in paper titles and then perform manual filtering. We also gradually include relevant papers missed by keyword search and papers from other venues encountered by following reference links throughout the surveying process.[2] The distribution of AL-related papers in the ACL Anthology over the past twenty years is shown in Figure 1, which also includes rough counts of works concerning neural models by searching for the word *"neural"* in titles. The overall trend is interesting. There is a peak around the years of 2009 and 2010, while the counts drop and fluctuate during the mid-2010s, which corresponds to the time when neural models became prominent in NLP. We observe a renewed interest in AL research in recent years, which is

[1]The descriptions in this survey are mostly brief to provide more comprehensive coverage in a compact way. We hope that this work can serve as an index for corresponding works.

[2]Appendix C describes more details of the process.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6166 - 6190
December 7-11, 2022 ©2022 Association for Computational Linguistics

Algorithm 1 A typical active learning procedure.

Input: An unlabeled data pool $\mathcal{U}$.
Output: The final labeled dataset $\mathcal{L}$ and trained model $\mathcal{M}$.

$\mathcal{L}, \mathcal{U} \leftarrow$ seed($\mathcal{U}$) ▷ **Start** (§5.1)
$\mathcal{M} \leftarrow$ train($\mathcal{L}, \mathcal{U}$) ▷ **Model Learning** (§4)
while not stop_criterion() **do** ▷ **Stop** (§5.2)
    $\mathcal{I} \leftarrow$ query($\mathcal{M}, \mathcal{U}$) ▷ **Query** (§2, §3)
    $\mathcal{I}' \leftarrow$ annotate($\mathcal{I}$) ▷ **Annotate** (§3)
    $\mathcal{U} \leftarrow \mathcal{U} - \mathcal{I}$; $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{I}'$
    $\mathcal{M} \leftarrow$ train($\mathcal{L}, \mathcal{U}$) ▷ **Model Learning** (§4)
return $\mathcal{L}, \mathcal{M}_f$

primarily focused on deep active learning (Ren et al., 2021; Zhan et al., 2022).

1.1 Overview

We mainly examine the widely utilized pool-based scenario (Lewis and Gale, 1994), where a pool of unlabeled data is available and instances are drawn from the pool to be annotated. Algorithm 1 illustrates a typical AL procedure, which consists of a loop of instance selection with the current model and model training with updated annotations. The remainder of this survey is organized corresponding to the main steps in this procedure:

- In §2, we discuss the core aspect of AL: Query strategies, with a fine-grained categorization over informativeness (§2.1), representativeness (§2.2) and the combination of these two (§2.3).
- In §3, we cover the two additional important topics of querying and annotating for NLP tasks: AL for structured prediction tasks (§3.1) and the cost of annotation with AL (§3.2).
- In §4, we discuss model and learning: the query-successor model mismatch scenario (§4.1) and AL with advanced learning techniques (§4.2).
- In §5, we examine methods for starting (§5.1) and stopping (§5.2) AL.

In §6, we conclude with related and future directions. We also include representative AL works for various NLP tasks in Appendix A and some other aspects of AL for NLP in Appendix B.

2 Query Strategies

2.1 Informativeness

Informativeness-based query strategies mostly assign an informative measure to each unlabeled instance *individually*. The instance(s) with the highest measure will be selected.

2.1.1 Output Uncertainty

Uncertainty sampling (Lewis and Gale, 1994) is probably the simplest and the most commonly utilized query strategy. It prefers the most uncertain instances judged by the model outputs. For probabilistic models, entropy-based (Shannon, 1948), least-confidence (Culotta and McCallum, 2005) and margin-sampling (Scheffer et al., 2001; Schein and Ungar, 2007) are three typical uncertainty sampling strategies (Settles, 2009). Schröder et al. (2022) revisit some of these uncertainty-based strategies with Transformer-based models and provide empirical results for text classification. For non-probabilistic models, similar ideas can be utilized, such as selecting the instances that are close to the decision boundary in an SVM (Schohn and Cohn, 2000; Tong and Koller, 2001).

Another way to measure output uncertainty is to check the divergence of a model's predictions with respect to an instance's local region. If an instance is near the decision boundary, the model's outputs may be different within its local region. In this spirit, recent works examine different ways to check instances' **local divergence**, such as nearest-neighbour searches (Margatina et al., 2021), adversarial perturbation (Zhang et al., 2022b) and data augmentation (Jiang et al., 2020).

2.1.2 Disagreement

Uncertainty sampling usually considers the outputs of only one model. In contrast, **disagreement-based strategies** utilize multiple models and select the instances that are most disagreed among them. This is also a widely-adopted algorithm, of which the famous query-by-committee (QBC; Seung et al., 1992) is an example. The disagreement can be measured by vote entropy (Engelson and Dagan, 1996), KL-divergence (McCallum and Nigam, 1998) or variation ratio (Freeman, 1965).

To construct the model committee, one can train a group of distinct models. Moreover, taking a Bayesian perspective over the model parameters is also applicable (Houlsby et al., 2011). Especially with neural models, (Gal and Ghahramani, 2016) show that dropout could approximate inference and measure model uncertainty. This **deep Bayesian method** has been applied to AL for computer vision (CV) tasks (Gal et al., 2017) as well as various NLP tasks (Siddhant and Lipton, 2018; Shen et al., 2018; Shelmanov et al., 2021).

2.1.3 Gradient

Gradient information can be another signal for querying, with the motivation to choose the instances that would most strongly impact the model.

In this strategy, informativeness is usually measured by the norm of the gradients. Since we do not know the gold labels for unlabeled instances, the loss is usually calculated as the expectation over all labels. This leads to the strategy of **expected gradient length** (EGL), introduced by Settles et al. (2007) and later applied to sequence labeling (Settles and Craven, 2008) and speech recognition (Huang et al., 2016). Zhang et al. (2017) explore a variation for neural networks where only the gradients of word embeddings are considered and show its effectiveness for text classification.

2.1.4 Performance Prediction

Predicting performance can be another indicator for querying. Ideally, the selected instances should be the ones that most **reduce future errors** if labeled and added to the training set. This motivates the expected error reduction strategy (Roy and McCallum, 2001), which chooses instances that lead to the least expected error if added to retrain a model. This strategy can be computationally costly since retraining is needed for each candidate.

Recently, methods have been proposed to learn another model to select instances that lead to the fewest errors, usually measured on a held-out development set. Reinforcement learning and imitation learning have been utilized to train such policy models (Bachman et al., 2017; Fang et al., 2017; Liu et al., 2018a,b). This **learning-to-select strategy** may have some constraints. First, it requires labeled data (maybe from another domain) to train the policy. To mitigate this reliance, Vu et al. (2019) use the current task model as an imperfect annotator for AL simulations. Moreover, the learning signals may be unstable for complex tasks, as Koshorek et al. (2019) show for semantic tasks.

A similar and simpler idea is to select the most erroneous or ambiguous instances with regard to the current task model, which can also be done with another **performance-prediction model**. Yoo and Kweon (2019) directly train a smaller model to predict the instance losses for CV tasks, which have been also adopted for NLP (Cai et al., 2021; Shen et al., 2021). In a similar spirit, Wang et al. (2017) employ a neural model to judge the correctness of the model prediction for SRL and Brantley et al. (2020) learn a policy to decide whether expert querying is required for each state in sequence labeling. Inspired by data maps (Swayamdipta et al., 2020), Zhang and Plank (2021) train a model to select ambiguous instances whose average correctness over the training iterations is close to a predefined threshold. For machine translation (MT), special techniques can be utilized to seek erroneous instances, such as using a backward translator to check round-trip translations (Haffari et al., 2009; Zeng et al., 2019) or quality estimation (Logacheva and Specia, 2014a,b).

2.2 Representativeness

Only considering the informativeness of individual instances may have the drawback of sampling bias (Dasgupta, 2011; Prabhu et al., 2019) and the selection of outliers (Roy and McCallum, 2001; Karamcheti et al., 2021). Therefore, representativeness, which measures how instances correlate with each other, is another major factor to consider when designing AL query strategies.

2.2.1 Density

With the motivation to avoid outliers, density-based strategies prefer instances that are more **representative of the unlabeled set**. Selecting by n-gram or word counts (Ambati et al., 2010a; Zhao et al., 2020b) can be regarded as a simple way of density measurement. Generally, the common measurement is an instance's average similarity to all other instances (McCallum and Nigam, 1998; Settles and Craven, 2008). While it may be costly to calculate similarities of all instance pairs, considering only k-nearest neighbor instances has been proposed as an alternative option (Zhu et al., 2008c, 2009).

2.2.2 Discriminative[3]

Another direction is to select instances that are **different from already labeled instances**. Again, for NLP tasks, simple feature-based metrics can be utilized for this purpose by preferring instances with more unseen n-grams or out-of-vocabulary words (Eck et al., 2005; Bloodgood and Callison-Burch, 2010; Erdmann et al., 2019). Generally, similarity scores can also be utilized to select the instances that are less similar to the labeled set (Kim et al., 2006; Zhang et al., 2018; Zeng et al., 2019). Another interesting idea is to train a model to discriminate the labeled and unlabeled sets. Gissin and Shalev-Shwartz (2019) directly train a classifier for this purpose, while naturally adversarial training can be also adopted (Sinha et al., 2019; Deng et al., 2018). In domain adaptation scenarios, the same

[3]Some works also use the word "diversity," however we specifically preserve this word for batch-diversity in §2.2.3.

motivation leads to the usage of a domain separator to filter instances (Rai et al., 2010).

2.2.3 Batch Diversity

Ideally, only one most useful instance would be selected in each iteration. However, it is more efficient and practical to adopt **batch-mode AL** (Settles, 2009), where each time a batch of instances is selected. In this case, we need to consider the dissimilarities not only between selected instances and labeled ones but also within the selected batch.

To select a batch of diverse instances, there are two common approaches. 1) **Iterative selection** collects the batch in an iterative greedy way (Brinker, 2003; Shen et al., 2004). In each iteration, an instance is selected by comparing it with previously chosen instances to avoid redundancy. Some more advanced diversity-based criteria, like coreset (Geifman and El-Yaniv, 2017; Sener and Savarese, 2018) and determinantal point processes (Shi et al., 2021), can also be approximated in a similar way. 2) **Clustering-based** methods partition the unlabeled data into clusters and select instances among them (Tang et al., 2002; Xu et al., 2003; Shen et al., 2004; Nguyen and Smeulders, 2004; Zhdanov, 2019; Yu et al., 2022). Since the chosen instances come from different clusters, diversity can be achieved to some extent.

For the calculation of similarity, in addition to comparing the input features or intermediate neural representations, other methods are also investigated, such as utilizing model-based similarity (Hazra et al., 2021), gradients (Ash et al., 2020; Kim, 2020), and masked LM surprisal embeddings (Yuan et al., 2020).

2.3 Hybrid

There is no surprise that informativeness and representativeness can be combined for instance querying, leading to hybrid strategies. A **simple combination** can be used to merge multiple criteria into one. This can be achieved by a weighted sum (Kim et al., 2006; Chen et al., 2011) or multiplication (Settles and Craven, 2008; Zhu et al., 2008c).

There are several strategies to **naturally integrate** multiple criteria. Examples include (uncertainty) weighted clustering (Zhdanov, 2019), diverse gradient selection (Ash et al., 2020; Kim, 2020) where the gradients themselves contain uncertainty information (§2.1.3) and determinantal point processes (DPP) with quality-diversity decomposition (Shi et al., 2021).

Moreover, **multi-step querying**, which applies multiple criteria in series, is another natural hybrid method. For example, one can consider first filtering certain highly uncertain instances and then performing clustering to select a diverse batch from them (Xu et al., 2003; Shen et al., 2004; Mirroshandel et al., 2011). An alternative strategy of selecting the most uncertain instances per cluster has also been utilized (Tang et al., 2002).

Instead of statically merging into one query strategy, **dynamic combination** may better fit the AL learning process, since different strategies may excel at different AL phases. For example, at the start of AL, uncertainty sampling may be unreliable due to little labeled data, and representativeness-based methods could be preferable, whereas in later stages where we have enough data and target finer-grained decision boundaries, uncertainty may be a suitable strategy. DUAL (Donmez et al., 2007) is such a dynamic strategy that can switch from a density-based selector to an uncertainty-based one. Ambati et al. (2011b) further propose GraDUAL, which gradually switches strategies within a switching range. Wu et al. (2017) adopt a similar idea with a pre-defined monotonic function to control the combination weights.

3 Query and Annotation

3.1 AL for Structured Prediction

AL has been widely studied for classification tasks, while in NLP, many tasks involve *structured prediction*. In these tasks, the system needs to output a structured object consisting of a group of interdependent variables (Smith, 2011), such as a label sequence or a parse tree. Special care needs to be taken when querying and annotating for these more complex tasks (Thompson et al., 1999). One main decision is whether to annotate full structures for input instances (§3.1.1), or allow the annotation of only partial structures (§3.1.2).

3.1.1 Full-structure AL

First, if we regard the full output structure of an instance as a whole and perform query and annotation at the full-instance level, then AL for structured prediction tasks is not very different than for simpler classification tasks. Nevertheless, considering that the output space is usually exponentially large and infeasible to explicitly enumerate, querying may require further inspection.

Some **uncertainty sampling** strategies, such as entropy, need to consider the full output space. Instead of the infeasible explicit enumeration, dynamic-programming algorithms that are similar to the ones in decoding and inference processes can be utilized, such as algorithms for tree-entropy (Hwa, 2000, 2004) and sequence-entropy (Mann and McCallum, 2007; Settles and Craven, 2008).

Instead of considering the full output space, **top-k approximation** is a simpler alternative that takes k-best predicted structures as a proxy. This is also a frequently utilized method (Tang et al., 2002; Kim et al., 2006; Rocha and Sanchez, 2013).

For disagreement-based strategies, the measurement of **partial disagreement** may be required, since full-match can be too strict for structured objects. Fine-grained evaluation scores can be reasonable choices for this purpose, such as F1 score for sequence labeling (Ngai and Yarowsky, 2000).

Since longer instances usually have larger uncertainties and might be preferred, **length normalization** is a commonly-used heuristic to avoid this bias (Tang et al., 2002; Hwa, 2000, 2004; Shen et al., 2018). Yet, Settles and Craven (2008) argue that longer sequences should not be discouraged and may contain more information.

Instead of directly specifying the full utility of an instance, **aggregation** is also often utilized by gathering utilities of its sub-structures, usually along the factorization of the structured modeling. For example, the sequence uncertainty can be obtained by summing or averaging the uncertainties of all the tokens (Settles and Craven, 2008). Other aggregation methods are also applicable, such as weighted sum by word frequency (Ringger et al., 2007) or using only the most uncertain (least probable) one (Myers and Palmer, 2021; Liu et al., 2022).

3.1.2 Partial-structure AL

A structured object can be decomposed into smaller sub-structures with different training utilities. For example, in a dependency tree, functional relations are usually easier to judge while prepositional attachment links may be more informative for the learning purpose. This naturally leads to AL with partial structures, where querying and annotating can be performed at the sub-structure level.

Factorizing full structures into the **finest-grained sub-structures** and regarding them as the annotation units could be a natural choice. Typical examples include individual tokens for sequence labeling (Marcheggiani and Artières, 2014), word boundaries for segmentation (Neubig et al., 2011; Li et al., 2012b), syntactic-unit pairs for dependency parsing (Sassano and Kurohashi, 2010) and mention pairs for coreference (Gasperin, 2009; Miller et al., 2012; Sachan et al., 2015). The querying strategy for the sub-structures can be similar to the classification cases, though inferences are usually needed to calculate marginal probabilities. Moreover, if full structures are desired as annotation outputs, semi-supervised techniques such as self-training (§4.2) could be utilized to assign pseudo labels to the unannotated parts (Tomanek and Hahn, 2009b; Majidi and Crane, 2013).

At many times, choosing **larger sub-structures** is preferable, since partial annotation still needs the understanding of larger contexts and frequently jumping among different contexts may require more reading time (§3.2.1). Moreover, increasing the sampling granularity may mitigate the missed class effect, where certain classes may be overlooked (Tomanek et al., 2009). Typical examples of larger sub-structures include sub-sequences for sequence labeling (Shen et al., 2004; Chaudhary et al., 2019; Radmard et al., 2021), word-wise head edges for dependency parsing (Flannery and Mori, 2015; Li et al., 2016), neighborhood pools (Laws et al., 2012) or mention-wise anaphoric links (Li et al., 2020; Espeland et al., 2020) for coreference, and phrases for MT (Bloodgood and Callison-Burch, 2010; Miura et al., 2016; Hu and Neubig, 2021). In addition to increasing granularity, **grouping queries** can also help to make annotation easier, such as adopting a two-stage selection of choosing uncertain tokens from uncertain sentences (Mirroshandel and Nasr, 2011; Flannery and Mori, 2015) and selecting nearby instances in a row (Miller et al., 2012).

For AL with partial structures, **output modeling** is of particular interest since the model needs to learn from partial annotations. If directly using local discriminative models where each sub-structure is decided independently, learning with partial annotations is straightforward since the annotations are already complete to the models (Neubig et al., 2011; Flannery and Mori, 2015). For more complex models that consider interactions among output sub-structures, such as global models, special algorithms are required to learn from incomplete annotations (Scheffer et al., 2001; Wanvarie et al., 2011; Marcheggiani and Artières, 2014; Li et al., 2016). One advantage of these more complex models is the interaction of the partial labels and the remaining parts. For example, considering the **output constraints** for structured prediction tasks, combining the annotated parts and the constraints may reduce the output space of other parts and thus lower their uncertainties, leading to better queries (Roth and Small, 2006; Sassano and Kurohashi, 2010; Mirroshandel and Nasr, 2011). More generally, the annotation of one label can intermediately influence others with cheap re-inference, which can help batch-mode selection (Marcheggiani and Artières, 2014) and interactive correction (Culotta and McCallum, 2005).

In addition to classical structured-prediction tasks, classification tasks can also be cast as structured predictions with partial labeling. **Partial feedback** is an example that is adopted to make the annotating of classification tasks simpler, especially when there are a large number of target labels. For example, annotators may find it much easier to answer yes/no questions (Hu et al., 2019) or rule out negative classes (Lippincott and Van Durme, 2021) than to identify the correct one.

3.2 Annotation Cost

AL mainly aims to reduce real annotation cost and we discuss several important topics on this point.

3.2.1 Cost Measurement

Most AL works adopt simple measurements of unit cost, that is, assuming that annotating each instance requires the same cost. Nevertheless, the annotation efforts for different instances may vary (Settles et al., 2008). For example, longer sentences may cost more to annotate than shorter ones. Because of this, many works assume unit costs to tokens instead of sequences, which may still be inaccurate. Especially, AL tends to select difficult and ambiguous instances, which may require more annotation efforts (Hachey et al., 2005; Lynn et al., 2012). It is important to **properly measure annotation cost** since the measurement directly affects the evaluation of AL algorithms. The comparisons of query strategies may vary if adopting different cost measurement (Haertel et al., 2008a; Bloodgood and Callison-Burch, 2010; Chen et al., 2015).

Probably the best cost measurement is the actual **annotation time** (Baldridge and Palmer, 2009). Especially, when the cost comparisons are not that straightforward, such as comparing annotating data against writing rules (Ngai and Yarowsky, 2000) or partial against full annotations (§3.1; Flannery and Mori, 2015; Li et al., 2016, 2020), time-based evaluation is an ideal choice. This requires actual annotating exercises rather than simulations.

Since cost measurement can also be used for querying (§3.2.2), it would be helpful to be able to **predict the real cost** before annotating. This can be cast as a regression problem, for which several works learn a linear cost model based on input features (Settles et al., 2008; Ringger et al., 2008; Haertel et al., 2008a; Arora et al., 2009).

3.2.2 Cost-sensitive Querying

Given the goal of reducing actual cost, the querying strategies should also take it into consideration. That is, we want to select not only high-utility instances but also low-cost ones. A natural cost-sensitive querying strategy is **return-on-investment** (ROI; Haertel et al., 2008b; Settles et al., 2008; Donmez and Carbonell, 2008). In this strategy, instances with higher net benefit per unit cost are preferred, which is equivalent to dividing the original querying utility by cost measure. Tomanek and Hahn (2010) evaluate the effectiveness of ROI together with two other strategies, including constraining maximal cost budget per instance and weighted rank combination. Haertel et al. (2015) provide further analytic and empirical evaluation, showing that ROI can reduce total cost.

In real AL scenarios, things can be much more complex. For example, there can be multiple annotators with different expertise (Baldridge and Palmer, 2009; Huang et al., 2017; Cai et al., 2020), and the annotators may refuse to answer or make mistakes (Donmez and Carbonell, 2008). Being aware of these scenarios, Donmez and Carbonell (2008) propose **proactive learning** to jointly select the optimal oracle and instance. Li et al. (2017) further extend proactive learning to NER tasks.

3.2.3 Directly Reducing Cost

In addition to better query strategies, there are other ways of directly reducing annotation cost, such as computer-assisted annotation. In AL, models and annotators usually interact in an indirect way where models only query the instances to present to the annotators, while there could be closer interactions.

Pre-annotation is such an idea, where not only the raw data instances but also the model's best or top-k predictions are sent to the annotators to help them make decisions. If the model's predictions are reasonable, the annotators can simply select or make a few corrections to obtain the gold annotations rather than creating from scratch.

This method has been shown effective when combined with AL (Baldridge and Osborne, 2004; Vlachos, 2006; Ringger et al., 2008; Skeppstedt, 2013; Cañizares-Díaz et al., 2021). Post-editing for MT is also a typical example (Dara et al., 2014).

Moreover, the models could provide help at **real annotating time**. For example, Culotta and McCallum (2005) present an interactive AL system where the user's corrections can propagate to the model, which generates new predictions for the user to further refine. Interactive machine translation (IMT) adopts a similar idea, where the annotator corrects the first erroneous character, based on which the model reproduces the prediction. AL has also been combined with IMT to further reduce manual efforts (González-Rubio et al., 2012; Peris and Casacuberta, 2018; Gupta et al., 2021).

3.2.4 Wait Time

In AL iterations, the annotators may need to wait for the training and querying steps (Line 3 and 4 in Algorithm 1). This wait time may bring some hidden costs, thus more efficient querying and training would be preferable for faster turnarounds.

To speed up **querying**, sub-sampling is a simple method to deal with large unlabeled pools (Roy and McCallum, 2001; Ertekin et al., 2007; Tsvigun et al., 2022). For some querying strategies, precalculating and caching unchanging information can also help to speed up (Ashrafi Asli et al., 2020; Citovsky et al., 2021). In addition, approximation with k-nearest neighbours can also be utilized to calculate density (Zhu et al., 2009) or search for instances after adversarial attacks (Ru et al., 2020).

To reduce **training** time, a seemingly reasonable strategy is to apply incremental training across AL iterations, that is, continuing training previous models on the new instances. However, Ash and Adams (2020) show that this type of warm-start may lead to sub-optimal performance for neural models and many recent AL works usually train models from scratch (Hu et al., 2019; Ein-Dor et al., 2020). Another method is to use an efficient model for querying and a more powerful model for final training. However, this might lead to sub-optimal results, which will be discussed in §4.1.

Another idea to reduce wait time is to simply allow querying with **stale information**. Actually, batch-mode AL (§2.2.3) is such an example where instances in the same batch are queried with the same model. Haertel et al. (2010) propose parallel AL, which maintains separate loops of annotating, training, and scoring, and allows dynamic and parameterless instance selection at any time.

4 Model and Learning

4.1 Model Mismatch

While it is natural to adopt the same best-performing model throughout the AL process, there are cases where the query and final (successor) models can mismatch (Lewis and Catlett, 1994). Firstly, more efficient models are preferable for querying to reduce wait time (§3.2.4). Moreover, since data usually outlive models, re-using AL-base data to train another model would be desired (Baldridge and Osborne, 2004; Tomanek et al., 2007). Several works show that model mismatch may make the gains from AL be negligible or even negative (Baldridge and Osborne, 2004; Lowell et al., 2019; Shelmanov et al., 2021), which raises concerns about the utilization of AL in practice.

For efficiency purposes, distillation can be utilized to improve querying efficiency while keeping reasonable AL performance. Shelmanov et al. (2021) show that using a smaller distilled version of a pre-trained model for querying does not lead to too much performance drop. Tsvigun et al. (2022) combine this idea with pseudo-labeling and sub-sampling to further reduce computational cost. Similarly, Nguyen et al. (2022) keep a smaller proxy model for query and synchronize the proxy with the main model by distillation.

4.2 Learning

AL can be combined with other advanced learning techniques to further reduce required annotations.

Semi-supervised learning. Since AL usually assumes an unlabeled pool, semi-supervised learning can be a natural fit. Combining these two is not a new idea: (McCallum and Nigam, 1998) adopt the EM algorithm to estimate the outputs of unlabeled data and utilize them for learning. This type of self-training or pseudo-labeling technique is often utilized in AL (Tomanek and Hahn, 2009b; Majidi and Crane, 2013; Yu et al., 2022). With a similar motivation, (Dasgupta and Ng, 2009) use an unsupervised algorithm to identify the unambiguous instances to train an active learner. For the task of word alignment, which can be learned in an unsupervised manner, incorporating supervision with AL can bring further improvements in a data-efficient way (Ambati et al., 2010b,c).

Transfer learning. AL can be easily combined with transfer learning, another technique to reduce required annotations. Utilizing pre-trained models is already a good example (Ein-Dor et al., 2020; Yuan et al., 2020; Tamkin et al., 2022) and continual training (Gururangan et al., 2020) can also be applied (Hua and Wang, 2022; Margatina et al., 2022). Moreover, transductive learning is commonly combined with AL by transferring learning signals from different domains (Chan and Ng, 2007; Shi et al., 2008; Rai et al., 2010; Saha et al., 2011; Wu et al., 2017; Kasai et al., 2019; Yuan et al., 2022) or languages (Qian et al., 2014; Fang and Cohn, 2017; Fang et al., 2017; Chaudhary et al., 2019, 2021; Moniz et al., 2022). In addition to the task model, the model-based query policy (§2.1.4) is also often obtained with transfer learning.

Weak supervision. AL can also be combined with weakly supervised learning. Examples include learning from inputs and execution results for semantic parsing (Ni et al., 2020), labeling based on identical structure vectors for entity representations (Qian et al., 2020), learning from gazetteers and dictionaries for sequence labeling (Brantley et al., 2020) and interactively discovering labeling rules (Zhang et al., 2022a).

Data augmentation. Augmentation is also applicable in AL and has been explored with iterative back-translation (Zhao et al., 2020b), mixup for sequence labeling (Zhang et al., 2020) and phrase-to-sentence augmentation for MT (Hu and Neubig, 2021). As discussed in §2.1.1, augmentation can also be helpful for instance querying (Jiang et al., 2020; Zhang et al., 2022b). Another interesting scenario involving augmentation and AL is query synthesis, which directly generates instances to be annotated instead of selecting existing unlabeled ones. Though synthesizing texts is still a hard problem generally, there have been successful applications for simple classification tasks (Schumann and Rehbein, 2019; Quteineh et al., 2020).

5 Starting and Stopping AL

5.1 Starting AL

While there are cases where there are already enough labeled data to train a reasonable model and AL is utilized to provide further improvements (Bloodgood and Callison-Burch, 2010; Geifman and El-Yaniv, 2017), at many times we are facing the cold-start problem, where instances need to be selected without a reasonable model. Especially, how to select the seed data to start the AL process is an interesting question, which may greatly influence the performance in initial AL stages (Tomanek et al., 2009; Horbach and Palmer, 2016).

Random sampling is probably the most commonly utilized strategy, which is reasonable since it preserves the original data distribution. Some representativeness-based querying strategies (§2.2) can also be utilized, for example, selecting points near the clustering centroids is a way to obtain representative and diverse seeds (Kang et al., 2004; Zhu et al., 2008c; Hu et al., 2010). Moreover, some advanced learning techniques (§4.2) can also be helpful here, such as transfer learning (Wu et al., 2017) and unsupervised methods (Vlachos, 2006; Dasgupta and Ng, 2009). In addition, language model can be a useful tool, with which Dligach and Palmer (2011) select low-probability words in the context of word sense disambiguation and Yuan et al. (2020) choose cluster centers with surprisal embeddings by pre-trained contextualized LMs.

5.2 Stopping AL

When adopting AL in practice, it would be desirable to know the time to stop AL when the model performance is already near the upper limits, before running out of all the budgets. For this purpose, a stopping criterion is needed, which checks certain metrics satisfying certain conditions. There can be simple heuristics. For example, AL can be stopped when all unlabeled instances are no closer than any of the support vectors with an SVM (Schohn and Cohn, 2000; Ertekin et al., 2007) or no new n-grams remain in the unlabeled set for MT (Bloodgood and Callison-Burch, 2010). Nevertheless, these are specific to the underlying models or target tasks. For the design of a general stopping criterion, there are three main aspects to consider: *metric*, *dataset* and *condition*.

For the **metric**, measuring performance on a development set seems a natural option. However, the results would be unstable if this set is too small and it would be impractical to assume a large development set. Cross-validation on the training set also has problems since the labeled data by AL is usually biased. In this case, metrics from the query strategies can be utilized. Examples include uncertainty or confidence (Zhu and Hovy, 2007; Vlachos, 2008), disagreement (Tomanek et al., 2007; Tomanek and Hahn, 2008; Olsson and Tomanek, 2009), estimated performance (Laws and Schütze, 2008), expected error (Zhu et al., 2008a), confidence variation (Ghayoomi, 2010), as well as actual performance on the selected instances (Zhu and Hovy, 2007). Moreover, comparing the predictions between consecutive AL iterations is another reasonable option (Zhu et al., 2008b; Bloodgood and Vijay-Shanker, 2009a).

The **dataset** to calculate the stopping metric requires careful choosing. The results could be unstable if not adopting a proper set (Tomanek and Hahn, 2008). Many works suggest that a separate unlabeled dataset should be utilized (Tomanek and Hahn, 2008; Vlachos, 2008; Bloodgood and Vijay-Shanker, 2009a; Beatty et al., 2019; Kurlandski and Bloodgood, 2022). Since the stopping metrics usually do not rely on gold labels, this dataset could potentially be very large to provide more stable results, though wait time would be another factor to consider in this case (§3.2.4).

The **condition** to stop AL is usually comparing the metrics to a pre-defined threshold. Earlier works only look at the metric at the current iteration, for example, stopping if the uncertainty or the error is less than the threshold (Zhu and Hovy, 2007). In this case, the threshold is hard to specify since it relies on the model and the task. (Zhu et al., 2008b) cascade multiple stopping criteria to mitigate this reliance. A more stable option is to track the change of the metrics over several AL iterations, such as stopping when the confidence consistently drops (Vlachos, 2008), the changing rate flattens (Laws and Schütze, 2008) or the predictions stabilize across iterations (Bloodgood and Vijay-Shanker, 2009a; Bloodgood and Grothendieck, 2013).

Pullar-Strecker et al. (2021) provide an empirical comparison over common stopping criteria and would be a nice reference. Moreover, stopping AL can be closely related to performance prediction and early stopping. Especially, the latter can be of particular interest to AL since learning in early AL stages need to face the low-resource problem and how to perform early stopping may also require careful considerations.

6 Related Topics and Future Directions

6.1 Related Topics

There are many related topics that could be explored together with AL. Other data-efficient learning methods such as semi-supervised and transfer learning are naturally compatible with AL (§4.2). Curriculum learning (Bengio et al., 2009), which arranges training instances in a meaningful order, may also be integrated with AL (Platanios et al., 2019; Zhao et al., 2020a; Jafarpour et al., 2021). Uncertainty (Gawlikowski et al., 2021), outlier detection (Hodge and Austin, 2004) and performance prediction (Xia et al., 2020) can be related to instance querying. Crowdsourcing can be adopted to further reduce annotation cost (§B). Model efficiency (Menghani, 2021) would be crucial to reduce wait time (§3.2.4). AL is a typical type of human-in-the-loop framework (Wang et al., 2021), and it will be interesting to explore more human-computer interaction techniques in AL.

6.2 Future Directions

Complex tasks. AL is mostly adopted for simple classification, while there are many more complex tasks in NLP. For example, except for MT, generation tasks have been much less thoroughly explored with AL. Tasks with more complex inputs such as NLI and QA also require extra care when using AL; obtaining unlabeled data is already non-trivial. Nevertheless, preliminary work has shown that AL can be helpful for data collection for such tasks (Mussmann et al., 2020).

Beyond direct target labeling. In addition to directly annotating target labels, AL can also be utilized in other ways to help the target task, such as labeling features or rationales (Melville and Sindhwani, 2009; Druck et al., 2009; Sharma et al., 2015), annotating explanations (Liang et al., 2020), evaluation (Mohankumar and Khapra, 2022) and rule discovery (Zhang et al., 2022a).

AL in practice. Most AL works simulate annotations on an existing labeled dataset. Though this method is convenient for algorithm development, it ignores several challenges of applying AL in practice. As discussed in this survey, real annotation cost (§3.2.1), efficiency and wait time (§3.2.4), data reuse (§4.1) and starting and stopping (§5) are all important practical aspects which may not emerge in simulation. Moreover, since the AL process usually cannot be repeated multiple times, how to select the query strategy and other hyper-parameters remains a great challenge. It will be critical to address these issues to bring AL into practical use (Rehbein et al., 2010; Attenberg and Provost, 2011; Settles, 2011; Lowell et al., 2019) and make it more widely utilized (Tomanek and Olsson, 2009).

Limitations

There are several limitations of this work. First, we mainly focus on AL works in the context of NLP, while AL works in other fields may also present ideas that could be utilized for NLP tasks. For example, many querying strategies originally developed with CV tasks could be naturally adopted to applications in NLP (Ren et al., 2021). We encourage the readers to refer to other surveys mentioned in §1 for additional related AL works. Moreover, the descriptions in this survey are mostly brief in order to provide a more comprehensive coverage within page limits. We mainly present the works in meaningful structured groups rather than plainly describing them in unstructured sequences, and we hope that this work can serve as an index where more details can be found in corresponding works. Finally, this is a pure survey without any experiments or empirical results. It would be helpful to perform comparative experiments over different AL strategies, which could provide more meaningful guidance (Zhan et al., 2022). We leave this to future work.

References

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. 2014. Active learning: A survey. In *Data Classification*, pages 599–634. Chapman and Hall/CRC.

Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel, and Jaime Carbonell. 2011a. Active learning with multiple annotations for comparable data classification task. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 69–77, Portland, Oregon. Association for Computational Linguistics.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010a. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010b. Active learning-based elicitation for semi-supervised word alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 365–370, Uppsala, Sweden. Association for Computational Linguistics.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010c. Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 10–17, Los Angeles, California. Association for Computational Linguistics.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011b. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010a. Discriminative sample selection for statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 626–635, Cambridge, MA. Association for Computational Linguistics.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010b. A semi-supervised batch-mode active learning strategy for improved statistical machine translation. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 126–134, Uppsala, Sweden. Association for Computational Linguistics.

Shilpa Arora, Eric Nyberg, and Carolyn P. Rosé. 2009. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 18–26, Boulder, Colorado. Association for Computational Linguistics.

Jordan Ash and Ryan P Adams. 2020. On warm-starting neural network training. *Advances in Neural Information Processing Systems*, 33:3884–3894.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, Reza Fahmi, and Omid Momenzadeh. 2020. Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2855–2861, Marseille, France. European Language Resources Association.

Jordi Atserias, Giuseppe Attardi, Maria Simi, and Hugo Zaragoza. 2010. Active learning for building a corpus of questions for parsing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Josh Attenberg and Şeyda Ertekin. 2013. Class imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, pages 101–149.

Josh Attenberg and Foster Provost. 2011. Inactive learning? difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter*, 12(2):36–41.

Philip Bachman, Alessandro Sordoni, and Adam Trischler. 2017. Learning algorithms for active learning. In *International Conference on Machine Learning*, pages 301–310. PMLR.

Guirong Bai, Shizhu He, Kang Liu, Jun Zhao, and Zaiqing Nie. 2020. Pre-trained language model based active learning for sentence matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1495–1504, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jason Baldridge and Miles Osborne. 2003. Active learning for HPSG parse selection. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 17–24.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16, Barcelona, Spain. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Garrett Beatty, Ethan Kochis, and Michael Bloodgood. 2019. The use of unlabeled data versus labeled data for stopping active learning for text classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 287–294. IEEE.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 854–864, Uppsala, Sweden. Association for Computational Linguistics.

Michael Bloodgood and John Grothendieck. 2013. Analysis of stopping active learning based on stabilizing predictions. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 10–19, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Bloodgood and K. Vijay-Shanker. 2009a. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 39–47, Boulder, Colorado. Association for Computational Linguistics.

Michael Bloodgood and K. Vijay-Shanker. 2009b. Taking into account the differences between actively and passively acquired data: The case of active learning with support vector machines for imbalanced datasets. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 137–140, Boulder, Colorado. Association for Computational Linguistics.

Kianté Brantley, Amr Sharaf, and Hal Daumé III. 2020. Active imitation learning with noisy guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2093–2105, Online. Association for Computational Linguistics.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 59–66.

Tingting Cai, Zhiyuan Ma, Hong Zheng, and Yangming Zhou. 2021. Ne–lp: normalized entropy-and loss prediction-based sampling for active learning in chinese word segmentation on ehrs. *Neural Computing and Applications*, 33(19):12535–12549.

Tingting Cai, Yangming Zhou, and Hong Zheng. 2020. Cost-quality adaptive active learning for chinese clinical named entity recognition. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 528–533. IEEE.

Hian Cañizares-Díaz, Alejandro Piad-Morffis, Suilan Estevez-Velarde, Yoan Gutiérrez, Yudivián Almeida Cruz, Andres Montoyo, and Rafael Muñoz-Guillena. 2021. Active learning for assisted corpus construction: A case study in knowledge discovery from biomedical text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 216–225, Held Online. INCOMA Ltd.

Kai Cao, Xiang Li, Miao Fan, and Ralph Grishman. 2015. Improving event detection with active learning. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 72–77, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5164–5174, Hong Kong, China. Association for Computational Linguistics.

Chenhua Chen, Alexis Palmer, and Caroline Sporleder. 2011. Enhancing active learning for semantic role labeling via compressed dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 183–191, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jinying Chen, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 120–127, New York City, USA. Association for Computational Linguistics.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning*, 15(2):201–221.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Aswarth Abhilash Dara, Josef van Genabith, Qun Liu, John Judge, and Antonio Toral. 2014. Active learning for post-editing based incrementally retrained MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 185–189, Gothenburg, Sweden. Association for Computational Linguistics.

Sajib Dasgupta and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, Suntec, Singapore. Association for Computational Linguistics.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781.

Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. 2018. Adversarial active learning for sequences labeling and generation. In *IJCAI*, pages 4012–4018.

Dmitriy Dligach and Martha Palmer. 2011. Good seed makes a good crop: Accelerating active learning using language modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 6–10, Portland, Oregon, USA. Association for Computational Linguistics.

Pinar Donmez and Jaime G Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628.

Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. 2007. Dual strategy active learning. In *European Conference on Machine Learning*, pages 116–127. Springer.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 81–90, Singapore. Association for Computational Linguistics.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 43–48, Melbourne, Australia. Association for Computational Linguistics.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Sean P. Engelson and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Santa Cruz, California, USA. Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. 2007. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 127–136.

Nuno Escudeiro and Alípio Jorge. 2010. D-confidence: An active learning strategy which efficiently identifies small classes. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 18–26, Los Angeles, California. Association for Computational Linguistics.

Vebjørn Espeland, Beatrice Alex, and Benjamin Bach. 2020. Enhanced labelling in active learning for coreference resolution. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–121, Barcelona, Spain (online). Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark. Association for Computational Linguistics.

Meng Fang, Jie Yin, and Dacheng Tao. 2014. Active learning for crowdsourcing using knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Daniel Flannery and Shinsuke Mori. 2015. Combining active learning and partial annotation for domain adaptation of a Japanese dependency parser. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 11–19, Bilbao, Spain. Association for Computational Linguistics.

Linton C Freeman. 1965. *Elementary applied statistics: for students in behavioral science*. New York: Wiley.

Lisheng Fu and Ralph Grishman. 2013. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 692–698, Nagoya, Japan. Asian Federation of Natural Language Processing.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1998. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Caroline Gasperin. 2009. Active learning for anaphora resolution. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 1–8, Boulder, Colorado. Association for Computational Linguistics.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*.

Masood Ghayoomi. 2010. Using variance as a stopping criterion for active learning of frame assignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *arXiv preprint arXiv:1907.06347*.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, Avignon, France. Association for Computational Linguistics.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kamal Gupta, Dhanvanth Boppana, Rejwanul Haque, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Investigating active learning in interactive neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 10–22, Virtual. Association for Machine Translation in the Americas.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Hossein Hadian and Hossein Sameti. 2014. Active learning in noisy conditions for spoken language understanding. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1081–1090, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Robbie Haertel, Paul Felt, Eric K. Ringger, and Kevin Seppi. 2010. Parallel active learning: Eliminating wait time with minimal staleness. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 33–41, Los Angeles, California. Association for Computational Linguistics.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and Peter McClanahan. 2008a. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 65–68, Columbus, Ohio. Association for Computational Linguistics.

Robbie Haertel, Eric Ringger, Kevin Seppi, and Paul Felt. 2015. An analytic and empirical evaluation of return-on-investment-based active learning. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 11–20, Denver, Colorado, USA. Association for Computational Linguistics.

Robbie A Haertel, Kevin D Seppi, Eric K Ringger, and James L Carroll. 2008b. Return on investment for active learning. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 72.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado. Association for Computational Linguistics.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 181–189, Suntec, Singapore. Association for Computational Linguistics.

Rishi Hazra, Parag Dutta, Shubham Gupta, Mohammed Abdul Qaathir, and Ambedkar Dukkipati. 2021. Active^2 learning: Actively reducing redundancies in active learning methods for sequence tagging and machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1982–1995, Online. Association for Computational Linguistics.

Rui He, Shan He, and Ke Tang. 2021. Multi-domain active learning: A comparative study. *arXiv preprint arXiv:2106.13516*.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*.

Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126.

Andrea Horbach and Alexis Palmer. 2016. Investigating active learning for short-answer scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 301–311, San Diego, CA. 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*.

Junjie Hu and Graham Neubig. 2021. Phrase-level active learning for neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1087–1099, Online. Association for Computational Linguistics.

Peiyun Hu, Zack Lipton, Anima Anandkumar, and Deva Ramanan. 2019. Active learning with partial feedback. In *International Conference on Learning Representations*.

Rong Hu, Brian Mac Namee, and Sarah Jane Delany. 2010. Off to a good start: Using clustering to select the initial training set in active learning. In *Twenty-Third International FLAIRS Conference*.

Xinyu Hua and Lu Wang. 2022. Efficient argument structure extraction with transfer learning and active learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 423–437, Dublin, Ireland. Association for Computational Linguistics.

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. 2016. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*.

Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. 2017. Cost-effective active learning from diverse labelers. In *IJCAI*, pages 1879–1885.

Rebecca Hwa. 2000. Sample selection for statistical grammar induction. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 45–52, Hong Kong, China. Association for Computational Linguistics.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3):253–276.

Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra. 2018. Multi-task active learning for neural semantic role labeling on low resource conversational corpus. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 43–50, Melbourne. Association for Computational Linguistics.

Makoto Imamura, Yasuhiro Takayama, Nobuhiro Kaji, Masashi Toyoda, and Masaru Kitsuregawa. 2009. A combination of active learning and semi-supervised learning starting with positive and unlabeled examples for word sense disambiguation: An empirical study on Japanese web search query. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 61–64, Suntec, Singapore. Association for Computational Linguistics.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active curriculum learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, Online. Association for Computational Linguistics.

Zhuoren Jiang, Zhe Gao, Yu Duan, Yangyang Kang, Changlong Sun, Qiong Zhang, and Xiaozhong Liu. 2020. Camouflaged Chinese spam content detection with semi-supervised generative active learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3080–3085, Online. Association for Computational Linguistics.

Jaeho Kang, Kwang Ryel Ryu, and Hyuk-Chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 384–388. Springer.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Seokhwan Kim, Yu Song, Kyungduk Kim, Jeong-Won Cha, and Gary Geunbae Lee. 2006. MMR-based active machine learning for bio named entity recognition. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 69–72, New York City, USA. Association for Computational Linguistics.

Yekyung Kim. 2020. Deep active learning for sequence labeling based on diversity and uncertainty in gradient. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 1–8, Suzhou, China. Association for Computational Linguistics.

Omri Koshorek, Gabriel Stanovsky, Yichu Zhou, Vivek Srikumar, and Jonathan Berant. 2019. On the limits of learning to actively learn semantic representations. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 452–462, Hong Kong, China. Association for Computational Linguistics.

Luke Kurlandski and Michael Bloodgood. 2022. Impact of stop sets on stopping active learning for text classification. *arXiv preprint arXiv:2201.05460*.

Florian Laws, Florian Heimerl, and Hinrich Schütze. 2012. Active learning for coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–512, Montréal, Canada. Association for Computational Linguistics.

Florian Laws, Christian Scheible, and Hinrich Schütze. 2011. Active learning with Amazon Mechanical Turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Florian Laws and Hinrich Schütze. 2008. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 465–472, Manchester, UK. Coling 2008 Organizing Committee.

Meisin Lee, Lay-Ki Soon, Eu Gene Siew, and Ly Fie Sugianto. 2022. CrudeOilNews: An annotated crude oil news corpus for event extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 465–479, Marseille, France. European Language Resources Association.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8320–8331, Online. Association for Computational Linguistics.

Maolin Li, Nhung Nguyen, and Sophia Ananiadou. 2017. Proactive learning for named entity recognition. In *BioNLP 2017*, pages 117–125, Vancouver, Canada,. Association for Computational Linguistics.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012a. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148, Jeju Island, Korea. Association for Computational Linguistics.

Shoushan Li, Guodong Zhou, and Chu-Ren Huang. 2012b. Active learning for Chinese word segmentation. In *Proceedings of COLING 2012: Posters*, pages 683–692, Mumbai, India. The COLING 2012 Organizing Committee.

Zhenghua Li, Min Zhang, Yue Zhang, Zhanyi Liu, Wenliang Chen, Hua Wu, and Haifeng Wang. 2016. Active learning for dependency parsing with partial annotation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 344–354, Berlin, Germany. Association for Computational Linguistics.

Weixin Liang, James Zou, and Zhou Yu. 2020. ALICE: Active learning with contrastive natural language explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4380–4391, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Dong-Ho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. 2019. AlpacaTag: An active learning-based crowd annotation framework for sequence tagging. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 58–63, Florence, Italy. Association for Computational Linguistics.

Thomas Lippincott and Ben Van Durme. 2021. Active learning and negative evidence for language identification. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 47–51, Online. Association for Computational Linguistics.

Bing Liu, Harrisen Scells, Guido Zuccon, Wen Hua, and Genghong Zhao. 2021. ActiveEA: Active learning for neural entity alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3374, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018a. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883, Melbourne, Australia. Association for Computational Linguistics.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018b. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 334–344, Brussels, Belgium. Association for Computational Linguistics.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. Ltp: A new active learning strategy for crf-based named entity recognition. *Neural Processing Letters*, pages 1–22.

Varvara Logacheva and Lucia Specia. 2014a. Confidence-based active learning methods for machine translation. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 78–83, Gothenburg, Sweden. Association for Computational Linguistics.

Varvara Logacheva and Lucia Specia. 2014b. A quality-based active sample selection strategy for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2690–2695, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shayne Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew Frank, Nikhil Ramesh, and Chris DuBois. 2022. Active learning over multiple domains in natural language tasks. *arXiv preprint arXiv:2202.00254*.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, Hong Kong, China. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Uí Dhonnchadha. 2012. Active learning and the Irish treebank. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 23–32, Dunedin, New Zealand.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden. Association for Computational Linguistics.

Saeed Majidi and Gregory Crane. 2013. Active learning for dependency parsing by a committee of parsers. In *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, pages 98–105, Nara, Japan. Assocation for Computational Linguistics.

Cyrielle Mallart, Michel Le Nouy, Guillaume Gravier, and Pascale Sébillot. 2021. Active learning for interactive relation extraction in a French newspaper's articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 886–894, Held Online. INCOMA Ltd.

Gideon Mann and Andrew McCallum. 2007. Efficient computation of entropy gradient for semi-supervised conditional random fields. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 109–112, Rochester, New York. Association for Computational Linguistics.

Diego Marcheggiani and Thierry Artières. 2014. An experimental comparison of active learning strategies for partially labeled sequences. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 898–906, Doha, Qatar. Association for Computational Linguistics.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Héctor Martínez Alonso, Barbara Plank, Anders Johannsen, and Anders Søgaard. 2015. Active learning for sense annotation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 245–249, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Andrew McCallum and Kamal Nigam. 1998. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 350–358.

Prem Melville and Vikas Sindhwani. 2009. Active dual supervision: Reducing the cost of annotating examples and features. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 49–57, Boulder, Colorado. Association for Computational Linguistics.

Vânia Mendonça, Ricardo Rei, Luisa Coheur, and Alberto Sardinha. 2022. Onception: Active learning with expert advice for real world machine translation. *arXiv preprint arXiv:2203.04507*.

Gaurav Menghani. 2021. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *arXiv preprint arXiv:2106.08962*.

Timothy Miller, Dmitriy Dligach, and Guergana Savova. 2012. Active learning for coreference resolution. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 73–81, Montréal, Canada. Association for Computational Linguistics.

Seyed Abolghasem Mirroshandel, Gholamreza Ghassem-Sani, and Alexis Nasr. 2011. Active learning strategies for support vector machines, application to temporal relation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 56–64, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Seyed Abolghasem Mirroshandel and Alexis Nasr. 2011. Active learning for dependency parsing using partially annotated sentences. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 140–149, Dublin, Ireland. Association for Computational Linguistics.

Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. 2016. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29, San Diego, California. Association for Computational Linguistics.

Akash Kumar Mohankumar and Mitesh Khapra. 2022. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8761–8781, Dublin, Ireland. Association for Computational Linguistics.

Joel Moniz, Barun Patra, and Matthew Gormley. 2022. On efficiently acquiring annotations for multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–85, Dublin, Ireland. Association for Computational Linguistics.

Ali Mottaghi, Prathusha K Sarma, Xavier Amatriain, Serena Yeung, and Anitha Kannan. 2020. Medical symptom recognition from patient text: An active learning approach for long-tailed multilabel distributions. *arXiv preprint arXiv:2011.06874*.

Stephen Mussmann, Robin Jia, and Percy Liang. 2020. On the Importance of Adaptive Data Collection for Extremely Imbalanced Pairwise Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3400–3413, Online. Association for Computational Linguistics.

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.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Grace Ngai and David Yarowsky. 2000. Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 117–125, Hong Kong. Association for Computational Linguistics.

Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79.

Minh Van Nguyen, Nghia Ngo, Bonan Min, and Thien Nguyen. 2022. FAMIE: A fast active learning framework for multilingual information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 131–139, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ansong Ni, Pengcheng Yin, and Graham Neubig. 2020. Merging weak and active supervision for semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8536–8543.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

Fredrik Olsson and Katrin Tomanek. 2009. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 138–146, Boulder, Colorado. Association for Computational Linguistics.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

Stanislav Peshterliev, John Kearney, Abhyuday Jagannatha, Imre Kiss, and Spyros Matsoukas. 2019. Active learning for new domains in natural language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 90–96, Minneapolis, Minnesota. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4058–4068, Hong Kong, China. Association for Computational Linguistics.

Zac Pullar-Strecker, Katharina Dost, Eibe Frank, and Jörg Wicker. 2021. Hitting the target: Stopping active learning at the cost-based optimum. *arXiv preprint arXiv:2110.03802*.

Kun Qian, Poornima Chozhiyath Raman, Yunyao Li, and Lucian Popa. 2020. Learning structured representations of entity names using Active Learning and weak supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6376–6383, Online. Association for Computational Linguistics.

Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. Bilingual active learning for relation classification via pseudo parallel corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Baltimore, Maryland. Association for Computational Linguistics.

Husam Quteineh, Spyridon Samothrakis, and Richard Sutcliffe. 2020. Textual data augmentation for efficient active learning on tiny datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7400–7410, Online. Association for Computational Linguistics.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence based deep active learning for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4310–4321, Online. Association for Computational Linguistics.

Piyush Rai, Avishek Saha, Hal Daumé, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, Los Angeles, California. Association for Computational Linguistics.

Ines Rehbein and Josef Ruppenhofer. 2011. Evaluating the impact of coder errors on active learning. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 43–51, Portland, Oregon, USA. Association for Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, and Alexis Palmer. 2010. Bringing active learning to life. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 949–957, Beijing, China. Coling 2010 Organizing Committee.

Roi Reichart and Ari Rappoport. 2009. Sample selection for statistical parsers: Cognitively driven algorithms and evaluation measures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 3–11, Boulder, Colorado. Association for Computational Linguistics.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*, pages 861–869, Columbus, Ohio. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40.

Eric Ringger, Marc Carmen, Robbie Haertel, Kevin Seppi, Deryle Lonsdale, Peter McClanahan, James Carroll, and Noel Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, pages 101–108, Prague, Czech Republic. Association for Computational Linguistics.

Martha-Alicia Rocha and Joan-Andreu Sanchez. 2013. Towards the supervised machine translation: Real word alignments and translations in a multi-task active learning process. In *Proceedings of Machine Translation Summit XIV: Posters*, Nice, France.

Dan Roth and Kevin Small. 2006. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer.

Dan Roth and Kevin Small. 2008. Active learning for pipeline models. In *AAAI*, pages 683–688.

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models. *arXiv preprint arXiv:2208.05379*.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 441–448.

Dongyu Ru, Jiangtao Feng, Lin Qiu, Hao Zhou, Mingxuan Wang, Weinan Zhang, Yong Yu, and Lei Li. 2020. Active sentence learning by adversarial uncertainty sampling in discrete space. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4908–4917, Online. Association for Computational Linguistics.

Mrinmaya Sachan, Eduard Hovy, and Eric P Xing. 2015. An active learning approach to coreference resolution. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. 2011. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer.

Manabu Sassano. 2002. An empirical study of active learning with support vector machines forJapanese word segmentation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Manabu Sassano and Sadao Kurohashi. 2010. Using smaller constituents rather than sentences in active learning for Japanese dependency parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 356–365, Uppsala, Sweden. Association for Computational Linguistics.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer.

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

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*, pages 839–846.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Raphael Schumann and Ines Rehbein. 2019. Active learning via membership query synthesis for semi-supervised sentence classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 472–481, Hong Kong, China. Association for Computational Linguistics.

Priyanka Sen and Emine Yilmaz. 2020. Uncertainty and traffic-aware active learning for semantic parsing. In *Proceedings of the First Workshop on Interactive and Executable Semantic Parsing*, pages 12–17, Online. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Seungmin Seo, Donghyun Kim, Youbin Ahn, and Kyong-Ho Lee. 2022. Active learning on pre-trained language model with task-independent triplet loss. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Burr Settles. 2009. Active learning literature survey.

Burr Settles. 2011. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1.

Burr Settles, Mark Craven, and Soumya Ray. 2007. Multiple-instance active learning. *Advances in neural information processing systems*, 20.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active learning with rationales for text classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 441–451, Denver, Colorado. Association for Computational Linguistics.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

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 of the Association for Computational Linguistics (ACL-04)*, pages 589–596, Barcelona, Spain.

Shirong Shen, Zhen Li, and Guilin Qi. 2021. Active learning for event extraction with memory-based loss prediction model. *arXiv preprint arXiv:2112.03073*.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. 2008. Actively transfer domain knowledge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 342–357. Springer.

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.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981.

Maria Skeppstedt. 2013. Annotating named entities in clinical text by combining pre-annotation and active learning. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 74–80, Sofia, Bulgaria. Association for Computational Linguistics.

Noah A Smith. 2011. Linguistic structure prediction. *Synthesis lectures on human language technologies*, 4(2):1–274.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022. Active learning helps pretrained models learn the intended task. *arXiv preprint arXiv:2204.08491*.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 120–127, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cynthia A Thompson, Mary Elaine Califf, and Raymond J Mooney. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 406–414.

Katrin Tomanek and Udo Hahn. 2008. Approximating learning curves for active-learning-driven annotation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Katrin Tomanek and Udo Hahn. 2009a. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, pages 105–112.

Katrin Tomanek and Udo Hahn. 2009b. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, Suntec, Singapore. Association for Computational Linguistics.

Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

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

Katrin Tomanek and Fredrik Olsson. 2009. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 45–48, Boulder, Colorado. Association for Computational Linguistics.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 486–495, Prague, Czech Republic. Association for Computational Linguistics.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Akim Tsvigun, Artem Shelmanov, Gleb Kuzmin, Leonid Sanochkin, Daniil Larionov, Gleb Gusev, Manvel Avetisian, and Leonid Zhukov. 2022. Towards computationally feasible deep active learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1198–1218, Seattle, United States. Association for Computational Linguistics.

Andreas Vlachos. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

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

Thuy-Trang Vu, Ming Liu, Dinh Phung, and Gholamreza Haffari. 2019. Learning how to active learn by dreaming. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4091–4101, Florence, Italy. Association for Computational Linguistics.

Chenguang Wang, Laura Chiticariu, and Yunyao Li. 2017. Active learning for black-box semantic role labeling with neural factors. In *IJCAI*.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Dittaya Wanvarie, Hiroya Takamura, and Manabu Okumura. 2011. Active learning with subsequence sampling strategy for sequence labeling tasks. *Information and Media Technologies*, 6(3):680–700.

Fangzhao Wu, Yongfeng Huang, and Jun Yan. 2017. Active sentiment domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1711, Vancouver, Canada. Association for Computational Linguistics.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2013. Online active learning for cost sensitive domain adaptation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 1–9, Sofia, Bulgaria. Association for Computational Linguistics.

Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. 2003. Representative sampling for text classification using support vector machines. In *European conference on information retrieval*, pages 393–407. Springer.

Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. 2011. Active learning from crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1161–1168.

Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, Hong Kong, China. Association for Computational Linguistics.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. 2022. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pei Zhang, Xueying Xu, and Deyi Xiong. 2018. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 153–158. IEEE.

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022a. Prompt-based rule discovery and boosting for interactive weakly-supervised learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 745–758, Dublin, Ireland. Association for Computational Linguistics.

Rongzhi Zhang, Yue Yu, and Chao Zhang. 2020. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8566–8579, Online. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022b. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association* *for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Ye Zhang, Matthew Lease, and Byron Wallace. 2017. Active discriminative text representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Mingjun Zhao, Haijiang Wu, Di Niu, and Xiaoli Wang. 2020a. Reinforced curriculum learning on pre-trained neural machine translation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 9652–9659.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 21–29, Gothenburg, Sweden. Association for Computational Linguistics.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020b. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Yunpeng Zhao, Mattia Prosperi, Tianchen Lyu, Yi Guo, Le Zhou, and Jiang Bian. 2020c. Integrating crowdsourcing and active learning for classification of work-life events from tweets. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 333–344. Springer.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

Zhong Zhou and Alex Waibel. 2021. Active learning for massively parallel translation of constrained text into low resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 32–43, Virtual. Association for Machine Translation in the Americas.

Hua Zhu, Wu Ye, Sihan Luo, and Xidong Zhang. 2020. A multitask active learning framework for natural language understanding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4900–4914, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingbo Zhu and Eduard Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 783–790, Prague, Czech Republic. Association for Computational Linguistics.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008a. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. 2008b. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1129–1136, Manchester, UK. Coling 2008 Organizing Committee.

Jingbo Zhu, Huizhen Wang, Benjamin K Tsou, and Matthew Ma. 2009. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6):1323–1331.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008c. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee.

A Tasks

In this section, we list representative works for different NLP tasks. According to the output structures, the tasks are further categorized into four groups: classification, sequence labeling, complex structured prediction, and generation.

Classification denotes the tasks whose output consists of only one variable. Text classification that assigns a target label to an input text sequence is a typical example. Pairwise classification and word-level classification are also commonly seen in NLP.

- **Text classification:** Please refer to the paper table mentioned in (§C) for related works. We do not list them here since there are too many.
- **Pairwise classification:** (Grießhaber et al., 2020; Bai et al., 2020; Mussmann et al., 2020)
- **Word sense disambiguation (WSD):** (Fujii et al., 1998; Chen et al., 2006; Chan and Ng, 2007; Zhu and Hovy, 2007; Zhu et al., 2008c; Imamura et al., 2009; Martínez Alonso et al., 2015)

Sequence labeling is probably the most commonly seen structured prediction task in NLP. It aims to predict a sequence of labels, among which there may be interactions and constraints.

- **Part-of-speech (POS):** (Engelson and Dagan, 1996; Ringger et al., 2007; Haertel et al., 2008a; Marcheggiani and Artières, 2014; Fang and Cohn, 2017; Brantley et al., 2020; Chaudhary et al., 2021)
- **(Named) entity recognition (NER/ER):** (Shen et al., 2004; Culotta and McCallum, 2005; Kim et al., 2006; Settles and Craven, 2008; Tomanek and Hahn, 2009b; Marcheggiani and Artières, 2014; Chen et al., 2015; Li et al., 2017; Shen et al., 2018; Siddhant and Lipton, 2018; Erdmann et al., 2019; Chaudhary et al., 2019; Brantley et al., 2020; Hazra et al., 2021; Shelmanov et al., 2021; Radmard et al., 2021)
- **Segmentation:** (Ngai and Yarowsky, 2000; Sassano, 2002; Neubig et al., 2011; Li et al., 2012b; Marcheggiani and Artières, 2014; Cai et al., 2021)
- **Natural language understanding (NLU):** (Hadian and Sameti, 2014; Deng et al., 2018; Peshterliev et al., 2019; Zhu et al., 2020)

Complex structure prediction in this work denotes the structure prediction tasks that are more complex than sequence labeling, and have *explicit* connections (alignments) between inputs and outputs. They usually aim to extract relational structures among input elements.

- **Parsing:** (Hwa, 2000; Tang et al., 2002; Baldridge and Osborne, 2003, 2004; Hwa, 2004; Reichart and Rappoport, 2009; Sassano and Kurohashi, 2010; Atserias et al., 2010; Mirroshandel and Nasr, 2011; Majidi and Crane, 2013; Flannery and Mori, 2015; Li et al., 2016; Shi et al., 2021)
- **Semantic role labeling (SRL):** (Roth and Small, 2006; Wang et al., 2017; Ikhwantri et al., 2018; Siddhant and Lipton, 2018; Koshorek et al., 2019; Myers and Palmer, 2021)
- **Coreference:** (Gasperin, 2009; Miller et al., 2012; Laws et al., 2012; Zhao and Ng, 2014; Sachan et al., 2015; Li et al., 2020; Espeland et al., 2020; Yuan et al., 2022)
- **Relation-related:** (Roth and Small, 2008; Bloodgood and Vijay-Shanker, 2009b; Mirroshandel et al., 2011; Fu and Grishman, 2013; Cañizares-Díaz et al., 2021; Mallart et al., 2021; Seo et al., 2022; Zhang et al., 2022a)
- **Event-related:** (Cao et al., 2015; Shen et al., 2021; Lee et al., 2022)
- **Word alignment:** (Ambati et al., 2010b,c; Rocha and Sanchez, 2013)
- **Entity alignment/resolution:** (Kasai et al., 2019; Liu et al., 2021)

Generation refers to the tasks that aim to generate a sequence of tokens. We differentiate them from plain structured prediction tasks since there are usually no explicit alignments between input and output sub-parts in the supervision and such alignments are usually implicitly modeled, especially in recent sequence-to-sequence neural models. MT is a typical generation task, where we further separate traditional statistical machine translation (SMT) and recent neural machine translation (NMT). We also include semantic parsing here, since recent works usually cast it as a sequence-to-sequence generation task.

- **SMT:** (Eck et al., 2005; Haffari et al., 2009; Haffari and Sarkar, 2009; Ananthakrishnan et al., 2010b; Bloodgood and Callison-Burch, 2010; Ambati et al., 2010a; Ananthakrishnan et al., 2010a; González-Rubio et al., 2012; Rocha and

Sanchez, 2013; Logacheva and Specia, 2014a,b; Miura et al., 2016)
- **NMT:** (Peris and Casacuberta, 2018; Liu et al., 2018b; Zhang et al., 2018; Zeng et al., 2019; Zhao et al., 2020b; Hu and Neubig, 2021; Gupta et al., 2021; Zhou and Waibel, 2021; Hazra et al., 2021; Mendonça et al., 2022)
- **Semantic parsing:** (Duong et al., 2018; Ni et al., 2020; Sen and Yilmaz, 2020)
- **Others:** (Mairesse et al., 2010; Deng et al., 2018)

B Other Aspects

We describe some other aspects that are frequently seen when applying AL to NLP.

Crowdsourcing and Noise. Crowdsourcing is another way to reduce annotation costs by including non-expert annotations (Snow et al., 2008). Naturally, AL and crowdsourcing may also be combined with the hope to further reduce cost (Ambati et al., 2010a; Laws et al., 2011; Yan et al., 2011; Fang et al., 2014; Zhao et al., 2020c). One specific factor to consider in this case is the noises in the crowdsourced data, since noisy data may have a negative impact on the effectiveness of AL (Rehbein and Ruppenhofer, 2011). Cost-sensitive querying strategies (§3.2.2) can be utilized to select both annotators and instances by estimating labelers' reliability (Yan et al., 2011; Fang et al., 2014). Requiring multiple annotations per instance and then consolidating is also applicable (Laws et al., 2011). Lin et al. (2019) provide a framework that enables automatic crowd consolidation for AL on the tasks of sequence labeling.

Multiple Targets. In many cases, we may want to consider multiple targets rather than only one, for example, annotating instances in multiple domains (Xiao and Guo, 2013; He et al., 2021; Longpre et al., 2022) or multiple languages (Haffari and Sarkar, 2009; Qian et al., 2014; Moniz et al., 2022). Moreover, there may be multiple target tasks, where multi-task learning (MTL) can interact with AL (Reichart et al., 2008; Ambati et al., 2011a; Rocha and Sanchez, 2013; Ikhwantri et al., 2018; Zhu et al., 2020; Rotman and Reichart, 2022). In these scenarios with multiple targets, naturally, strategies that consider all the targets are usually more preferable. Reichart et al. (2008) show that a query strategy that considers all target tasks obtains the overall best performance for MTL. Moniz et al. (2022) suggest that joint learning across multiple languages using a single model outperforms other strategies such as equally dividing budgets or allocating only for a high-resource language and then performing the transfer.

Data Imbalance. Imbalance is a frequently occurring phenomenon in NLP and AL can have interesting interactions with it. On the one hand, as in plain learning scenarios, AL should take data imbalance into considerations, with modifications to the model (Bloodgood and Vijay-Shanker, 2009b), learning algorithm (Zhu and Hovy, 2007) and query strategies (Tomanek et al., 2009; Escudeiro and Jorge, 2010; Li et al., 2012a). On the other hand, AL can be utilized to address the data imbalance problem and build better data (Ertekin et al., 2007; Tomanek and Hahn, 2009a; Attenberg and Ertekin, 2013; Mottaghi et al., 2020; Mussmann et al., 2020).

C Surveying Process

In this section, we provide more details of our surveying process:

- For the ACL Anthology, we search for papers with the keyword "active" in titles (by grepping the "Full Anthology BibTeX file"[4]). There can be related papers that are missed from this simple keyword search, but as we read along the filtered list, we gradually include the notable missing ones.

- We also include papers outside the ACL Anthology. First, we look for papers by searching with the key phrase "active learning" on Arxiv (in the field of cs.CL, excluding those already appearing in ACL Anthology). Moreover, we also collect related works in other venues, such as AI/ML conferences and journals. For these venues, we do (can) not perform extensive searches due to high volume (and that many are unrelated to our focus on NLP). We mainly collect related papers in these adjacent venues by following the references from the papers already surveyed.

We also create a table for the related papers (with detailed categorizations), which can be found at this link: https://github.com/zzsfornlp/zmsp/blob/main/msp2/docs/al4nlp/readme.md.

[4] https://aclanthology.org/anthology.bib.gz