On the use of *Query by Committee* for Human-in-the-Loop Named Entity Recognition

Gabriel Corvino[1], Vitor Vasconcelos Oliveira[1], Angelo C. Mendes da Silva[2], Ricardo Marcondes Marcacini[2]

[1] Universidade de Brasília, Brazil
{gabriel.corvino, vasconcelos.oliveira}@aluno.unb.br
[2] Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, Brazil
{angelo.mendes, ricardo.marcacini}@usp.br

Abstract. Named Entity Recognition is a relevant task for extracting information from textual data. Traditional methods for training NER models assume that humans annotate entities manually, identifying entities in predefined categories. This strategy presents a great human effort, mainly in more specific application domains. To address these challenges, we consider Human in the Loop (HITL), which can be understood as a set of strategies to incorporate human knowledge and experience into machine learning, while accelerating model training. In this paper, we investigate a classic method called Query by Committee (QBC), which helps to select informative instances for data labeling. Traditionally, QBC selects instances with a high level of disagreement between different models of a committee. We present heuristics for QBC relaxation to also consider some level of agreement. We showed that taking advantage of some committee agreement for pre-labeling of instances is promising to speed up human feedback and increase the training set. Experimental results showed that our method is able to preserve the performance of models compared to traditional QBC, while reducing human labeling effort.

CCS Concepts: • **Computing methodologies** → *Natural language processing*.

Keywords: Human in the Loop, Active Learning, Ensemble, Named Entity Recognition

1. INTRODUCTION

Human-in-the-Loop (HITL) is a set of strategies to incorporate human knowledge and experience to increase the accuracy of a machine learning model while achieving a certain target accuracy for a model faster [Zanzotto 2019; Monarch 2021]. Recently, HITL has been successfully applied to handle data selection, labeling, and accelerating model training for NER tasks [Monarch 2021]. In particular, a crucial step in HITL is the data sampling technique for human interaction. Active Learning (AL) techniques are explored in this step, aiming to select the best subset of data [Ren et al. 2021]. In particular, AL are machine learning methods that interactively query humans or knowledge bases to label data, usually minimizing the number of queries according to some criteria. Many criteria have been proposed in the last decades [Aggarwal et al. 2014; Ren et al. 2021], such as uncertainty sampling [Lewis and Gale 1994], density sampling [Settles and Craven 2008], and query by committee sampling [Seung et al. 1992]. This last criterion is especially useful for scenarios involving human-in-the-loop, since humans can interact with different committee models (human-model interaction), and the models of a committee can interact with each other (model-model interaction).

We focus on Query by Committee (QBC), a method proposed by Seung, Opper, and Sompolinksy [Seung et al. 1992] for active learning of classification tasks. Initially, different models are trained from the same set of labeled data. Then, committee members can vote on each instance (unlabeled data) to define the entity's category. The basic idea of QBC is that the most informative instances are

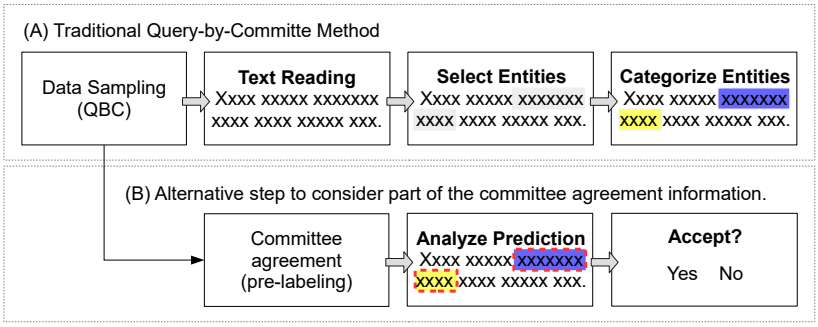

Fig. 1. Illustration of data labeling processes for NER tasks. In (A) represents the traditional labeling process in which the committee selects the entities and the human interaction categorizes them. In (B) illustrates a pre-labeling entity process based on committee agreement information, reducing the human effort to categorize the entity in a binary task.

those where most models disagree. The query returns such instances so that humans can incorporate their experience and knowledge. Previous studies have formally demonstrated that QBC generalizes learning by incorporating diversity into the training set from instances that cause maximum disagreement amongst the committee [Seung et al. 1992; Melville and Mooney 2004], i.e., querying unlabeled instances from controversial regions of the input space [Kumar and Gupta 2020].

This paper presents a relaxation of the QBC method for HITL scenarios, especially to reduce human effort in data labeling for Named Entity Recognition tasks. While the traditional QBC only focuses on disagreement, we argue that some committee agreement information is useful to support the human-based labeling of named entities. Figure 1 illustrates the proposed QBC relaxation in the context of HITL-NER. In Figure 1(A), we present the traditional labeling process, in which QBC selects an informative textual instance for HITL interaction. In this sense, humans perform a costly process of text analysis, involving reading the document, as well as identifying and categorizing the named entities. This labeled data is inserted into the training set for the next iteration of updating the NER models. Note that in this case, the agreement information is discarded and not used in the labeling interface. On the other hand, in Figure 1(B), we chose not to discard instances with some level of agreement and use them for pre-labeling the textual document. Users can approve or disapprove the labeled instance suggestion, i.e., a binary interaction that is much less costly in the HITL process. Some studies claim that binary responses are the best strategy for quality control and should be chosen wherever possible in the HITL process (see [Monarch 2021] for a discussion of interfaces for active learning involving HITL). We summarize below the main contributions of this paper:

QBC relaxation for HITL. Our method mostly explores disagreements in HITL-NER according to the original theoretical framework of QBC. However, we introduce a heuristic to include some instances that have a level of agreement to speed up the labeling process without reducing the accuracy of the trained NER models. We point out that incorporating model predictions into data labeling interfaces is a technique that has been explored for decades. In general, existing approaches are based on model prediction confidence. However, to the best of our knowledge, there are no previous studies investigating QBC in the context of HITL-NER, in which some level of agreement is exploited to pre-annotate data and reduce the cost of human feedback.

Early Reduction of Error Propagation. Some models may agree and be incorrect about the named entity. This situation is not directly handled by the original QBC. Such instances are also informative, and our method allows humans to correct the error in advance (e.g., disapproval of a prediction), thereby avoiding the propagation of this error to the next iterations.

Using Partially Correct Predictions. Small changes in an entity's words can be considered an error by NER models during the model evaluation, negatively impacting the traditional QBC disagreement strategy. Our method exploits partially correct predictions in QBC relaxation for HITL-NER, so that

humans can speed up data labeling with simple adjustments to the pre-labeled entity (e.g., simple addition or removal of a token from a pre-labeled entity).

We carried out an experimental evaluation on three real-world datasets involving entities about (1) laptop features, (2) restaurant aspects, and (3) journal text entities. We compared our method against the original QBC, considering a measure of human labeling effort in a HITL approach. Our method obtained promising results, as it preserved the accuracy of the NER models in relation to those trained via QBC sampling while reducing the human labeling effort by 12.6% on average.

2. QUERY BY COMMITTEE FOR HITL-NER

We instantiate the Human-in-the-Loop (HITL) framework[1] presented by [Song et al. 2021] for the Named Entity Recognition (NER) tasks, as illustrated in Figure 2. The contribution of this paper is in the sampling stage, in which we introduce a relaxation heuristic to the Query by Committee (QBC) method aiming to explore some agreement information to pre-label data and reduce human effort in the annotation stage.

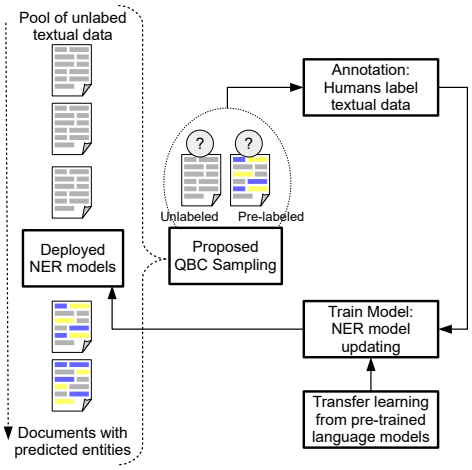

Fig. 2. A mental model of the human-in-the-loop process for named entity recognition. Adapted from [Monarch 2021].

Note that the HITL framework assumes a Transfer Learning stage using deep neural language models. This strategy allows starting the HITL process with less labeled data through some pre-trained models. In our method, we use transfer learning by fine-tuning different contextual language models. Combining multiple pre-trained models and transfer learning are useful techniques to mitigate perpetuating model bias, especially when there are small labeled data at the beginning of a HITL process.

Now, we formally introduce Query By Committee (QBC) and our proposed heuristics for QBC relaxation. Let $D_n = \{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$ be a training set with n textual instances, where x_i represents a sequence of tokens and y_i represents labeled information about token entities (e.g. IOB notation). We denote $H(M, D_n)$ as the entropy of the current state, in which models M were trained from training set D_n. Given a new training instance (x_{n+1}, y_{n+1}), we retrain the models from D_{n+1}, thus reaching the entropy $H(M, D_{n+1})$ for the new state. In fact, we aim to select an instance (or set of instances) that increases the information gain as defined in Equation 1,

$$I_{n+1} = H(M, D_n) - H(M, D_{n+1}) \tag{1}$$

[1]The original framework was proposed for general classification tasks, while ours was adapted specifically for NER tasks.

where if we select a D_{n+1} that reduces entropy in the new state (i.e., $H(M, D_{n+1}) < H(M, D_n)$), then there is an improvement in the learning process. Note that we are assuming that a low entropy value given M models is associated with the models' ability to correctly categorize new text entities. There is no consensus on the ideal number of models or the best strategy to estimate entropy, as it depends on the application and task. Our work uses classical idea proposed by [Dagan and Engelson 1995], which estimates entropy as $-\sum_t \frac{V(t,w)}{k} \log \frac{V(t,w)}{k}$, where $V(t, w)$ is the number of votes from a committee with k models that identified the word w with the tag t.

Calculating the information gain via Equation 1 can be computationally prohibitive. Thus, different heuristics are proposed to identify good candidates to increase the training set. In this sense, Query by Committee (QBC) uses the disagreement concept. If a given instance is classified differently by the vast majority of M models, then determining the correct label for that instance helps to reduce entropy. In practice, this instance is informative and a good candidate for human labeling. A theoretical analysis of the QBC was presented by [Seung et al. 1992], in which the basic idea is to minimize the version space, i.e., a set of hypotheses that are consistent according to the current training set [Settles 2009].

Although the disagreement strategy is promising for identifying informative instances, this strategy requires a greater human annotation effort. Without using pre-labeled information about the instance, users must label the text following the traditional process: reading the document, identifying and categorizing the named entities (see Figure 1(A)). We propose QBC disagreement relaxation heuristics to reduce such human annotation effort, according to Equation 2,

$$R_{n+1} = H(M, D_n) - H(M, D_{n+1}) - Q(x_{n+1}) \tag{2}$$

where $Q(x_{n+1})$ indicates the cost to annotate the new instance x_{n+1}; and R_{n+1} represents a score function (to be maximized) when inserting the new instance. Our QBC relaxation is given by considering a trade-off in information gain, $H(M, D_n) - H(M, D_{n+1})$, and annotation cost $Q(x_{n+1})$. Next, we present three heuristics to calculate $Q(x_{n+1})$:

(1) An instance that obtained some level of agreement between models has a α cost, as we can use predicted information for pre-labeling the document to be presented to the user (see Figure 1(B)), who can approve or reject. If approved, then the instance is added to the training set and can be useful to improve the training of other models that did not agree with the instance's prediction. If rejected, then the instance is added to the training set and is useful for minimizing error propagation and updating models.

(2) An instance that obtained some level of partial agreement in the predictions, i.e., models agree with part of the entity's tokens, has a β cost. In this case, the human annotation effort can also be reduced, as adjustments are made to the tokens pre-annotated by the models instead of a full annotation (i.e., reading the document, analyzing and categorizing entities). If the pre-annotation is incorrect, the cost for editing the pre-labeled data is not far from the cost of full annotation.

(3) An instance that has reached full agreement among all models is not considered for annotation, i.e., $Q(x_{n+1}) \to \infty$. This heuristic aims not to include redundant instances in the training set. We argue that the first heuristic is sufficient to identify early error propagation.

Defining the cost values α and β for heuristics 1 and 2, respectively, can be seen as an empirical task. If such values are too high, the trend of QBC relaxation is to select instances focusing mainly on reducing human labeling costs. However, this strategy will fail since QBC will not obtain quality instances for human labeling and model training. On the other hand, if these values are close to zero, the method will obtain results similar to the traditional QBC. In the next section, we present an experimental evaluation to discuss our method's practical aspects in more detail.

3. EXPERIMENTAL EVALUATION

We use three datasets for experimental evaluation. The Laptops and Restaurants datasets (English) were obtained from the Semeval 2015 [Pontiki et al. 2015] competition and adapted for named entity tasks considering an end-to-end aspect-based sentiment analysis. The named entities represent aspects of product reviews (aspects) and services (restaurants). The category/tag indicates the entity's sentiment (neutral, positive, and negative) in the text review. The Government gazettes dataset contains texts (Portuguese) from a Brazilian government newspaper. Named entities involve names of people, positions, and respective jobs for government agencies. Each dataset has predefined training and testing folds.

We used a committee of NER models based on five pre-trained neural language models. Each model was fine-tuned considering the initial training set, as well as the training data obtained via QBC sampling. For the English datasets (Laptops and Restaurants), we use the following pre-trained models: BERT (base-cased) [Devlin et al. 2019], RoBERTa (base) [Liu et al. 2019], DistilBERT (base-cased) [Sanh et al. 2019], XLNet (base-cased) [Yang et al. 2019], and AlBERT (base v2) [Lan et al. 2020]. For the Portuguese dataset (Government gazettes), we use the following pre-trained models: BERTimbau (base-cased) [Souza et al. 2020], DistilBERT-PT (base), NER-News-Portuguese (base-cased), Lener-BR (base-cased), and Wikineural-Multilingual [Tedeschi et al. 2021].

A NER model is obtained from a pre-trained neural language model via fine-tuning strategy [Sun et al. 2019], where we add a BI-LSTM layer for sequence labeling on top of the pre-trained model. Each NER model is trained (fine-tuning) with 10 epochs and a learning rate of 0.001. We simulate a Human-in-the-loop process according to the following steps: (1) Models are initially trained considering a small percentage p of the training set; (2) For each iteration of HTIL-NER, the QBC method must return a number of documents for human annotation. The query size used was: 125 for Laptops, 200 for Restaurants, and 250 for Government gazettes. We used five iterations for training models with HITL-NER.

3.1 Results and Discussion

We present and discuss the experimental results considering two aspects. The first is to show a performance overview (F1-Score) obtained by the NER methods from a HITL-NER process, comparing the traditional QBC and the QBC with the proposed relaxation heuristics. The second is to analyze the reduction of users' labeling effort when using some levels of agreement from the committee for pre-labeling instances during the HITL-NER.

Table I presents an overview of the F1-Score obtained for each NER model in each dataset. We present the F1-Score of the initial model and the final models obtained after five iterations of the HITL-NER. Note that both QBC methods were successful in querying instances that improved the initial model — an expected behavior since the initial model was trained with a reduced amount of labeled data. Furthermore, we found no statistically significant difference between the two approaches. In fact, another relevant observation is that the relaxation proposed in the QBC allowed for preserving the quality of the instance sampling since the F1-Score measures are similar to the traditional QBC.

On the other hand, our QBC allowed users' labeling effort reduction by 12.6% on average. Considering the evaluation through simulation of user interaction, we inferred the reduction of the labeling effort based on the relaxation heuristics proposed in the QBC. We use the value $\alpha = 0.75$, indicating that each correctly pre-labeled instance reduces the effort of annotating a particular instance by 75%. We use the value $\beta = 0.25$, i.e., partially correct pre-labeled instances reduce user effort by 25% on the annotation task. These measures are applied to all proposed QBC queries during the 5 iterations of the HITL-NER iterations, where the total labeling effort reduction is the sum of the reductions of each instance. Although they are arbitrarily defined values for our experimental evaluation, we derived these values from an analysis of the average time for users to answer questions in binary or

Table I. F1-Score measures considering the initial model and the respective final models trained through HITL-NER.

	Laptops		
Model	**Initial Model**	**QBC (Disagreement)**	**QBC (Ours)**
NER-AlBERT	0.49	0.58	0.57
NER-BERT	0.45	0.57	0.56
NER-DistilBERT	0.48	0.56	0.56
NER-RoBERTa	0.38	0.66	0.62
NER-XLNet	—	0.55	0.53
Labeling Effort Reduction	—	—	12%
	Restaurants		
Model	**Initial Model**	**QBC (Disagreement)**	**QBC (Ours)**
NER-AlBERT	0.54	0.68	0.70
NER-BERT	0.55	0.65	0.67
NER-DistilBERT	0.54	0.62	0.64
NER-RoBERTa	0.55	0.70	0.68
NER-XLNet	0.40	0.63	0.55
Labeling Effort Reduction	—	—	18%
	Government gazettes		
Model	**Initial Model**	**QBC (Disagreement)**	**QBC (Ours)**
NER-BERTimbau	0.80	0.90	0.87
NER-DistilBERT-PT	0.73	0.87	0.87
NER-LenerBR	0.80	0.86	0.87
NER-News	0.74	0.86	0.87
NER-Wikineural	0.73	0.86	0.85
Labeling Effort Reduction	—	—	8%

non-binary interfaces in crowdsourced environments [Alonso 2019].

To highlight the result of the proposed heuristics, we present in Table II the number of candidate instances for pre-labeling in HITL-NER in each iteration, considering the Restaurants dataset and the NER-AlBERT model. Each HITL iteration selects 200 instances. The first iteration already increases the training set and presents a significant F1-Score increase in relation to the initial model (iteration = 0). The disagreement column (%) indicates the percentage of disagreement of the query instances. The heuristics columns indicate the number of pre-labeled candidate instances for analysis with reduced user effort.

We observed that the first iterations allow selecting more candidate instances to reduce user effort. This behavior can be explained by the fact that the easiest instances to classify are initially selected by our QBC since several models agree on their entity categories. Throughout the HITL process, the remaining instances are the most difficult to classify and, consequently, the level of disagreement between the models increases. In this case, our method acts similarly to the traditional QBC.

Table II. Number of candidate instances for pre-labeling in HITL-NER using Heuristic #1 and #2, considering the Restaurants dataset and the NER-AlBERT model.

Iteration	**Disagreement (%)**	**F1-Score**	**Heuristic #1**	**Heuristic #2**
0	—	0.54	—	—
1	67.3%	0.64	21	35
2	74.6%	0.64	17	19
3	70.3%	0.65	14	39
4	85.6%	0.66	5	1
5	89.0%	0.70	3	4

4. RELATED WORK

Adopting a single strategy to select instances within an active learning process can result in a bias in model training by prioritizing instances with similar structures. As commonly used in classification

tasks, integrating multiple strategies with different selection criteria allows bias reduction in the selection process [Beluch et al. 2018], and creates heterogeneous subsets more similar to the data distribution. In this sense, [Seung et al. 1992] propose the approach Query by Committee (QBC), which discusses an iterative learning process where a committee is consulted at each iteration, resulting in feedback that allows measuring the relevance of each instance. QBC mainly focuses on selecting instances with the greatest disagreement between committee models. The central idea is to combine selection instances approaches with different preferences to support decision-making according to the committee [Zhao et al. 2006]. As a strategy to qualify the feedback, the approach called Human in the Loop (HITL) [Monarch 2021] has been explored in NER, which proposes the inclusion of a specialist user during the iterations and incorporating their experience in the NER training process model. However, due to the difficulty in scaling this process to large volumes of data, an important challenge for HITL is estimating human effort within active learning [Laws and Schütze 2008].

In our work, the integration of HITL in NER tasks is optimized by including a committee-based method responsible for measuring which instances are most relevant and require human review. As highlighted in [Wu et al. 2022], there is great interest in applications that use HITL to improve model performance through interventional model training, but the NER task is under-explored in the literature. In addition, we have incorporated an approach to early error propagation reduction using partially correct predictions, which can speed up the labeling text process.

In fact, our paper revisits Query by Committee in the HITL-NER context [Seung et al. 1992]. We argue that the Query by Committee (QBC) natural multiple-model strategy has theoretical advantages from the point of view of generalization robustness. In addition, intermediate results of outputs from various models can be aggregated to pre-label some instances. This pre-label information is useful for modifying NER's complex user annotation interface to a simple binary feedback interface, approving or rejecting the suggestion made by a subset of models.

5. CONCLUDING REMARKS

We argue that Human in the Loop for Named Entity Recognition (HITL-NER) is a naturally costly task, as it involves reading documents, annotating text excerpts, as well as identifying and categorizing entities. Our paper discusses promising alternatives to reduce this cost, revisiting Query by Committee as a promising strategy for HITL-NER. First, we reduce the bias of specific models by considering multiple NER models. Second, we apply heuristics aiming at QBC relaxation by exploring some level of agreement for pre-labeling documents. Experimental results on three datasets showed that our approach preserves the performance of the original QBC while reducing approximately 12.6% of human effort. More details about our method, source code and datasets are available at `https://github.com/rmarcacini/hitl-machine-learning`.

A limitation of our study is the use of a simulated strategy to evaluate the reduction of labeling effort. In this simulation, we do not consider that the annotator can make mistakes. In real scenarios, it is important to consider human annotators with different levels of knowledge and annotation errors. Another limitation of the study is how to estimate the reduction in human effort from labeling. There is no well-defined metric for NER, so we derive from other studies involving crowdsourcing with different interfaces (binary choice or not). Thus, directions for future work involve investigating our QBC method considering noisy labels and weak supervision, as well as an empirical study with humans estimating the real annotation time in each approach.

Acknowledgments: The authors would like to thank FAPDF project KnEDLe (grant 07/2019).

REFERENCES

Aggarwal, C. C., Kong, X., Gu, Q., Han, J., and Philip, S. Y. Active learning: A survey. In *Data Classification*. CRC Press, USA, pp. 599–634, 2014. Publisher Copyright: 2015 by Taylor & Francis Group LLC.

Alonso, O. Algorithms and techniques for quality control. In *The Practice of Crowdsourcing*. Springer, Cham, pp. 53–63, 2019.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE/CVF, Salt Lake City, Utah, EUA, pp. 9368–9377, 2018.

Dagan, I. and Engelson, S. P. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*. Morgan Kaufmann, San Francisco (CA), pp. 150–157, 1995.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, 2019.

Kumar, P. and Gupta, A. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology* 35 (4): 913–945, 2020.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In *8th International Conference on Learning Representations*. Open Review, Addis Ababa, Ethiopia, pp. 1–17, 2020.

Laws, F. and Schütze, H. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*. ACL, USA, pp. 465–472, 2008.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, London, UK, pp. 3–12, 1994.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Melville, P. and Mooney, R. J. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*. Association for Computing Machinery, New York, NY, USA, pp. 74, 2004.

Monarch, R. M. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, UK, 2021.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation*. Association for Computational Linguistics, Denver, Colorado, pp. 486–495, 2015.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. A survey of deep active learning. *ACM Computing Surveys (CSUR)* 54 (9): 1–40, 2021.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Settles, B. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Settles, B. and Craven, M. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*. Association for Computational Linguistics, USA, pp. 1070–1079, 2008.

Seung, H. S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Association for Computing Machinery, New York, NY, USA, pp. 287–294, 1992.

Song, B., Li, F., Liu, Y., and Zeng, X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics* 22 (6): 282, 2021.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*. Springer International Publishing, Cham, pp. 403–417, 2020.

Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*. Springer, Springer International Publishing, Cham, pp. 194–206, 2019.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 2521–2533, 2021.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* vol. 135, pp. 364–381, 2022.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* vol. 32, pp. 11, 2019.

Zanzotto, F. M. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* vol. 64, pp. 243–252, 2019.

Zhao, Y., Xu, C., and Cao, Y. Research on query-by-committee method of active learning and application. In *Advanced Data Mining and Applications*, X. Li, O. R. Zaïane, and Z. Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 985–991, 2006.