An active learning pipeline to automatically identify candidate terms for a CDSS ontology—measures, experiments, and performance

Shailesh Alluri, MS1, Keerthana Komatineni*, MS1, Rohan Goli*, MS1, Richard D. Boyce, PhD2, Nina Hubig, PhD1, Hua Min, PhD3, Yang Gong, PhD4, Dean F. Sittig, PhD4, David Robinson, MD5, Paul Biondich, MD6, Adam Wright, PhD7, Christian Nøhr, PhD8, Timothy Law, DO9, Arild Faxvaag, PhD10, Ronald Gimbel, PhD11, Lior Rennert, PhD11, Xia Jing#, PhD11

1School of Computing, College of Engineering, Computing and Applied Sciences, Clemson University, Clemson, SC, USA; 2Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; 3Department of Health Administration and Policy, College of Public Health, George Mason University, Fairfax, VA, USA; 4Department of Clinical and Health Informatics, McWilliams School of Biomedical Informatics, The University of Texas Health Sciences Center at Houston, Houston, TX, USA; 5Independent Consultant, Cumbria, UK; 6Department of Pediatrics, School of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA; 7Department of Medicine, School of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA; 8Department of Sustainability and Planning, The Technical Faculty of IT and Design, Danish Center for Health Informatics, Aalborg University, Aalborg, Denmark; 9Ohio Musculoskeletal and Neurologic Institute, Ohio University, Athens, OH, USA; 10Department of Neuromedicine and Movement Science, School of Medicine, Norwegian University of Science and Technology, Trondheim, Norway; 11Department of Public Health Sciences, College of Behavioral, Social and Health Sciences, Clemson University, Clemson, SC; *Coauthors contributed equally; #Corresponding author.

ABSTRACT:

Objective: To explore new strategies to make the document selection process more transparent, reproducible, and effective for the active learning process. The ultimate goal is to leverage active learning in identifying keyphrases to facilitate ontology development and construction, to streamline the process, and help with the long-term maintenance.
Methods: The active learning pipeline used a BILSTM-CRF model and over 2900 abstracts retrieved from PubMed relevant to clinical decision support systems. We started the model training with synthetic labeled abstracts, then used different strategies to select domain experts' annotated abstracts (gold standards). Random sampling was used as the baseline. Recall, F1 (beta = 1, 5, and 10) scores are used as measures to compare the performance of the active learning pipeline by different strategies.
Results: We tested four novel document-level uncertainty aggregation strategies—KPSum, KPAvg, DOCSum, and DOCAvg—that operate over standard token-level uncertainty scores such as Maximum Token Probability (MTP), Token Entropy (TE), and Margin. All strategies show significant improvement in early active learning cycles (θ_0 to θ_2) for recall and F1. The systematic evaluations show that KPSum (actual order) shows consistent improvement in both recall and F1 and KPSum (actual order) shows better results than the random sampling results.

The document order (actual versus reverse) does not seem to play a critical role across strategies in model learning and performance in our datasets, although in some strategies, actual order shows slightly more effective results. The weighted F1 (beta = 5 and 10) provided complementary results to raw recall and F1 (beta = 1).
Conclusion: While prior work on uncertainty sampling typically focuses on token-level uncertainty metrics within generic NER tasks, our work advances this line of research by introducing a higher-level abstraction: document-level uncertainty aggregation. With a human-in-the-loop Active Learning pipeline, it can effectively prioritize high-impact documents, improve early-cycle recall, and reduce annotation effort. Our results show promise in automating part of ontology construction and maintenance work, i.e., monitoring and screening new publications to identify candidate keyphrases. However, future work needs to improve the model performance to make it usable in real-world operations.

INTRODUCTION:

Clinical Decision Support Systems (CDSS) have been a cornerstone in modern healthcare, aiding medical practitioners by offering data-driven recommendations to improve patient care [Ref 1]. Ontology is an enabling technology of the Semantic Web technologies and plays a critical role in information interoperability [Ref 2]. A CDSS ontology defines the CDSS entities and their relationships, and can be used in multiple ways, including to develop reusable CDSS rules. The ontologies and terminologies are crucial in standardizing clinical vocabulary and ensuring interoperability across different healthcare systems [Ref 3]. However, building and maintaining ontologies is labor-intensive, time-consuming, and requires domain expertise. In the long term, the ontology requires continuous updates to include new discoveries. As medical knowledge rapidly evolves, the task of maintaining ontologies manually becomes increasingly challenging. Without consistent maintenance after construction, the outdated ontologies could provide outdated domain information. Although description logics [Ref 4] and dedicated teams can overcome some of the challenges, automating some of the steps can provide significant relief in long-term ontology maintenance.

This paper proposes an active learning-based pipeline that automates the identification of key phrases from unstructured CDSS publications. The key phrases can include a single word, a noun, a verb, or a phrase. The key phrases can be used as candidates to be evaluated by experts before being added to the CDSS ontology. Our approach leverages deep learning models to systematically select and highlight important terms. The model is refined by incorporating iterative human domain expert (HDE) feedback. This reduces the reliance on solely manual selection while ensuring the ontology remains accurate and current. The approach is particularly important in long-term CDSS ontology maintenance. Our pipeline prioritizes instances where the model is uncertain or where consensus is mostly needed, enabling targeted human review to resolve ambiguities effectively.

Statement of Significance

(1) Problem or Issue

Domain ontology development and maintenance are labor-intensive and time-consuming processes, primarily conducted by HDE manually.

(2) What is Already Known

Active learning approaches are used to identify key phrases in scientific literature. However, most of the measures of uncertainty are at the token level, which can make the contexts less focused and less coherent during model training/learning, considering the document is the unit for annotation.

(3) What this Paper Adds

We proposed, tested, and compared new approaches to calculate uncertainty measures for each document: **KPSum**, **KPAvg**, **DOCSum**, and **DOCAvg.** This enables more effective, informative, and transparent document selection for annotation in **keyphrase identification.** We conducted a set of systematic experiments to compare uncertainty strategies (e.g., **MTP-KPSum-Actual**, **TE-DocAvg-Reverse**, etc.) against baseline random sampling via evaluating **F1**, **recall**, and **β-weighted F1 scores** over active learning cycles for a CDSS ontology.

(4)Who would benefit from the new knowledge in this paper

Our approach had the potential to make the ontology development and maintenance more efficient and manageable over time. The work will benefit HDE directly.

The remainder of this paper is structured as follows**:** Section 2 presents relevant work; Section 3 presents the deep learning model architecture with an active learning loop, and our novel confidence score calculation methods; Section 4 outlines the experimental setup and results; Section 5 interprets the primary results with a brief discussion; Section 6 is the conclusion.

Section 2: Literature

Identifying key phrases from textual data is a critical task in natural language processing (NLP), forming the foundation for downstream applications such as ontology development, information retrieval, and text summarization. Traditional methods, including rule-based and statistical approaches like Term Frequency-Inverse Document Frequency and Rapid Automatic Keyword Extraction [Ref 6], have been widely adopted for key phrase extraction. However, these methods often fail to generalize across domains, prompting the adoption of machine learning-based models, such as conditional random fields (CRFs) [Ref 7] and sequence labeling models, for named entity recognition (NER) and key phrase identification.

Models such as Bidirectional Long Short-Term Memory Network - Conditional Random Fields (BiLSTM-CRF) [Ref 8] and transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT) [Ref 9] have set new benchmarks in NER tasks. Notably, the work by Alzaidy et al. utilized [Ref 10] a BiLSTM-CRF model for key phrase identification, demonstrating its ability to effectively leverage hierarchical context and transfer learning techniques for domain-specific applications.

Active learning (AL) is a subfield of machine learning that aims to achieve high model performance with minimal labeled data by strategically selecting the most informative samples for annotation [Ref 11]. Settles (2009) [Ref 12] provides a comprehensive overview of active

learning strategies, highlighting uncertainty sampling, query-by-committee, and expected error reduction as key techniques. Shen et al. (2017) [Ref 13] employed uncertainty sampling with deep learning models, effectively reducing the labeling effort required for text classification. Similarly, Siddhant and Lipton (2018) [Ref 14] explored active learning in low-resource NER tasks, demonstrating the benefits of selectively annotating high-value samples. Caufield, et al also explored using large language models (GPT-3.5 and GPT-4) with structured prompts to populate ontologies, to leverage the most advanced technology to achieve similar goals [Ref 15].

Leveraging active learning for key phrase identification has gained attention and demonstrated good efficiency and accuracy in recent years. Radmard et al. (2021) [Ref 16] utilized active learning to iteratively refine NER models, focusing on identifying ambiguous or novel terms in unstructured text. Margatina et al. (2021) [Ref 17] extended active learning techniques to transformer models like BERT, achieving significant performance gains with reduced annotation costs.

Our work differs from the above-mentioned work in two ways. First, prior studies rely solely on a small manually labeled seed set; our approach begins with a synthetically labeled training set, enabling the model to adapt to real-world patterns progressively. Second, we introduce novel uncertainty aggregation strategies designed specifically for document-level sampling. Traditional token-level uncertainty scores are inadequate in our setting, where the sampling unit is an entire document.

Section 3: METHODOLOGY

3.1 Active Learning Overview

A typical supervised learning setting requires a large amount of labeled data to train a model effectively. However, in many real-world applications, obtaining labeled data is expensive and time-consuming due to the need for domain expertise. AL addresses this challenge by iteratively and strategically selecting the data points that the model finds most uncertain and presenting them for human annotation, which reduces the annotation effort while maximizing the model's learning efficiency [Ref 18].
An AL cycle refers to the learning process between two stages, e.g., θ_x to θ_{x+1}, where θ_x represents the model state after x cycles of AL. Each cycle involves selecting the most informative (i.e., the most uncertain) samples, obtaining human annotations, and retraining the model with the newly human-labeled data (Figure 1). This iterative approach ensures that the model continuously improves by focusing on the most uncertain samples in each cycle.

Our AL cycle operates as follows (Figure 1):

- At stage θ_x, a BiLSTM-CRF model is used to classify tokens into five possible tags: B-KP (Beginning of a key phrase), I-KP (Inside a key phrase), O-KP (Outside a key phrase), Start (Start of the sequence), and Stop (End of the sequence). The model

assigns probabilities to each token, indicating the likelihood of belonging to one of these categories. These probabilities serve as the foundation for uncertainty measures and confidence scores computation later.

- After the confidence scores calculation (sections 3.3.1 and 3.3.2), the documents are sorted based on these scores. The sorting ensures that the most informative (i.e., uncertain) documents are prioritized for HDE annotation.
- After annotation, the HDE-labeled documents are added to the updated training data and removed from the AL data pool. This step is essential to avoid overfitting, as repeatedly training on the same document could reinforce that document rather than improving the model's generalizability.
- Once the newly labeled documents are incorporated into the updated training dataset, the BiLSTM-CRF model is retrained using the updated training dataset. This newly trained model, denoted as θ_{x+1}, then undergoes the same AL selection process in the next cycle. This iterative process continues until a stopping criterion (such as performance saturation) is met.

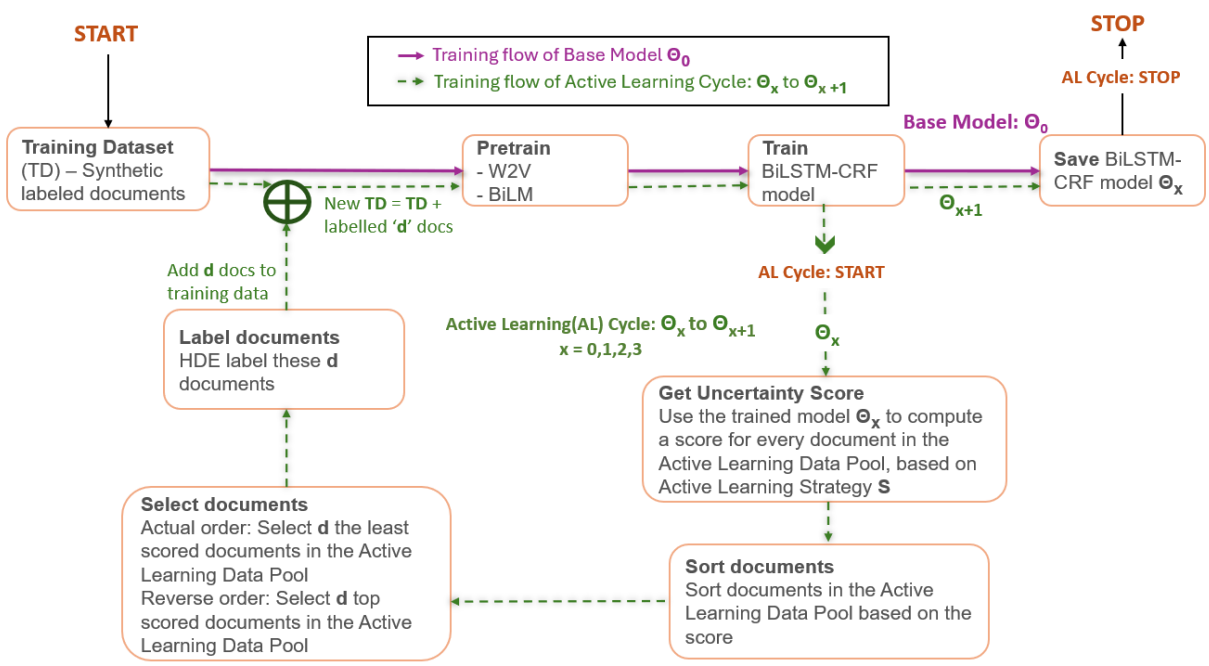


Figure 1: Active Learning Cycle for automatic key phrase identification

3.3 Active Learning Framework

Our AL framework begins with a BiLSTM-CRF model, i.e., the foundational model (θ_0) for key phrase identification. θ_0 is trained on 500 synthetically-labeled CDSS documents and our prior publication detailed the process [Ref 10]. Then, a pool-based AL strategy, as outlined by Settles et al. [Ref 12], was used to iteratively enhance the model by selecting and labeling the most

informative documents. This process allows the model to progressively improve its performance by leveraging uncertainty sampling techniques, focusing on the most uncertain data points to maximize learning efficiency. Figure 2 illustrates the entire AL lifecycle, showing the iterative model learning and improvement process from θ_0 to θ_4.

- θ_0: The initial model (θ_0) is trained on 500 synthetically labeled documents [Ref 10].
- θ_1: After the first AL cycle, the model is retrained on 500 synthetic documents + 5 selected documents reviewed by HDE.
- θ_2: The training dataset expands further, now including 500 synthetic documents + 10 HDE-reviewed documents (5 from θ_1 + 5 new documents).
- θ_3: The model is retrained on 500 synthetic documents + 15 HDE-reviewed documents (10 from θ_1 and θ_2 + 5 new documents).
- θ_4: The final cycle includes 500 synthetic documents + 20 HDE-reviewed documents (15 from θ_1, θ_2, and θ_3 + 5 new documents), continuing the refinement process.

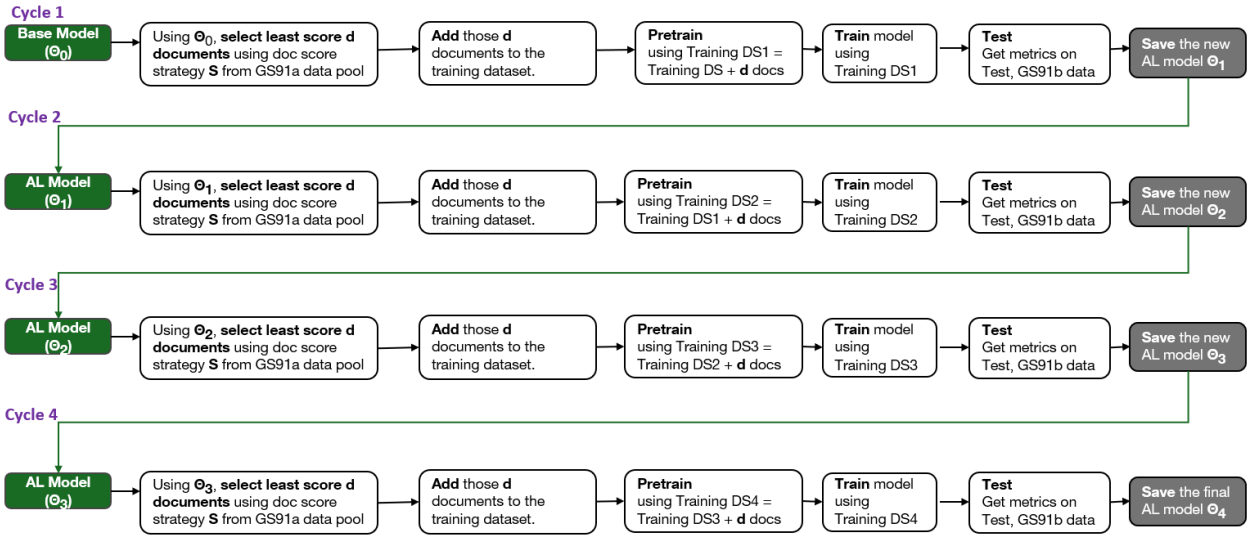


Figure 2: Active Learning life cycle examples from θ_0 to θ_4 with elaborated steps

3.3.1 Uncertainty Sampling Strategies: To identify sample documents for HDE annotation, we utilize well-established uncertainty metrics such as Minimum Token Probability (MTP), Margin, and Token Entropy. These metrics help identify tokens or sequences for which the model's predictions exhibit the lowest confidence.

Minimum Token Probability (MTP) [Ref 5]

The MTP assesses uncertainty by considering the highest probability assigned to any label for a token. A lower highest probability indicates greater uncertainty, as the model lacks confidence in any single prediction.

$$\phi^{MTP}(x) = 1 - \min_i \max_j P(y_i = j \mid x_i; A)$$

Where j ranges over all possible token tags. A is a parameter called the transition matrix. X_i is the i-th token in the sentence, and y_i is the tag that the i-th token can assume.

Token Entropy [Ref 19]

The **Token Entropy** measures uncertainty by calculating the entropy of the probability distribution across all possible tags for each token. Greater entropy indicates less information available to the model to make a confident prediction for a given token. By selecting tokens with the greatest entropy for AL, the system becomes efficient at gaining information for future predictions. Token Entropy is well-suited for key phrase identification, as it provides a probability-based uncertainty, allowing for targeting uncertain areas.

$$\phi_{TE} = -\sum_{j=1}^{M} P(y_i = j \mid x_i; A) \log P(y_i = j \mid x_i; A)$$

Where j ranges over all possible token tags. A is a parameter called a transfer matrix. X_i is the token. Y_i is the tag that the ith token in the sentence can assume M is the number of tags, in our case it is 5.

Margin Method [Ref 20]

The **Margin Method** calculates uncertainty by examining the difference between the probabilities of the two most likely predicted labels for a token. A smaller margin reflects greater uncertainty, as the model finds it harder to differentiate between the top predictions.
While the method still involves computing the full probability distribution, it only utilizes the top two probabilities, making it computationally more efficient than methods like Token Entropy, which require contributions from the entire distribution. The Margin Method is effective for identifying instances near prediction boundaries, where the model's confidence is weakest.

$$\phi_M(x) = -\left(P(y_1^* \mid x; A) - P(y_2^* \mid x; A)\right)$$

Where A is a parameter called a transfer matrix. $y^*{}_1$ and $y^*{}_2$ are the most and the second most likely tags for the token x, respectively.

3.3.2 Document-Level Aggregation: We introduce novel methods—KPSum, KPAvg, DOCSum, and DOCAvg—to aggregate token-level confidence scores into document-level confidence scores. Gupta used similar measures, ChowSum and ChowAverage [Ref 21], both of which aggregate uncertainty at the sentence level. Our approach differs in two ways. First, document-level aggregation better aligns with our AL task, where entire documents are labeled. Second, we propose KPSum and KPAvg—that aggregate uncertainty only over tokens likely to belong to keyphrases, enabling more targeted and efficient sampling. This keyphrase-centric perspective is absent in prior work, distinguishing our method from others.

DocSum sums the model uncertainty (measured by methods such as Token Entropy or Margin, explained above) across all tokens in a document, i.e., a cumulative uncertainty of all tokens. **DocSum** identifies documents with the highest overall uncertainty for HDE review. However, **DocSum** tends to favor longer documents, as they naturally contain more tokens.

DocAvg calculates the arithmetic mean uncertainty per token in the document. This method ensures that documents with high mean uncertainty (rather than just long documents with more tokens) are selected, balancing document length and uncertainty.

KPSum (Key Phrase Sum) sums the uncertainty scores of all identified key phrases (instead of all tokens) within a document. It prioritizes documents based on the collective uncertainty of the predicted key phrases. However, KPSum can be biased toward documents with more key phrases, i.e., more predicted key phrases will contribute more to the score.

KPAvg (Key Phrase Avg) takes the arithmetic average uncertainty of the predicted key phrases per document, regardless of the number of predicted key phrases in a document.

Both KPSum and KPAvg focus on the most ambiguous key phrases predicted by the BiLSTM-CRF model. These methods prioritize key phrase tokens over all tokens. Therefore, KPSum and KPAvg ensure that both the model and HDE attention are directed toward documents containing the most uncertain key phrases. A summary table about different strategies is included in the Appendix (Table A1).

Document Sorting and Selection: After computing the document-level confidence scores, all documents in the Active Learning data pool are sorted in descending order of their uncertainty for HDE to review.

Human Domain Expert Review: The most uncertain documents, as identified through the confidence scores, are reviewed and annotated by the HDE.

Retraining the Model: Once the documents are labeled by HDE, they will be added to the training dataset and removed from the AL data pool. The BiLSTM-CRF model is then retrained using the updated training dataset. This process allows the model to learn from the continuing updated correct labels and gradually improve its key phrase identification performance.

4. Experiments

This section outlines the experimental design used to evaluate the various AL strategies for key phrase identification, including the datasets, the metrics, the document selection order applied across AL cycles, and the results. Table 1 summarizes the systematic evaluation of different measures, strategies, and their combinations conducted during the experiments.

4.1 Experiment preparation

4.1.1 Document Order in Active Learning Cycles

The order to select documents during AL cycles plays a significant role in shaping the model learning trajectory. We evaluate two strategies:

1. **Actual Order**: Documents are prioritized based on their uncertainty (e.g., calculated by Token Entropy, Margin, or Maximum Token Probability). This strategy mimics a real-world scenario where the most uncertain documents are presented first for HDE annotation.
2. **Reverse Order**: Documents are presented in reverse order of their uncertainty. This serves as a control to evaluate whether prioritizing high-uncertainty documents truly enhances model performance compared to random or low-priority sampling.

4.1.2 Experiment Strategies: In each experiment, we test a specific strategy, which is a combination of the uncertainty measure (e.g., Margin, MTP, Token Entropy), document aggregator (e.g., DOCSum, DOCAvg, KPSum, KPAvg), and document order (e.g., Actual, Reverse). These combinations collectively determine how documents are prioritized for HDE annotation during AL cycles. Recall and F1 improvement were used to compare the model's performance.

Table 1. All the active learning strategies (i.e., combinations of uncertainty measure and strategy) experimented with in this study

Active Learning Strategies Evaluated in the Experiments				
Uncertainty Measure/ Document Aggregator	**DOCSum**	**DOCAvg**	**KPSum**	**KPAvg**
MTP	A:NA R: NA	A: S6 R: S5	A: S8 R: S7	A: S10 R: S9
MARGIN	A:NA R: NA	A: S2 R: S1	A: S3 R: S4	A:NA R: NA
TOKEN ENTROPY	A:NA R: NA	A: S12 R: S11	A: S14 R: S13	A:NA R: NA

Note: NA: Not Available; A: Actual order of uncertainty scores; R: Reverse order of uncertainty scores; Sx : Strategy x.

The DOCSum-based strategies are not included in the final performance comparison, as they consistently favored longer documents. However, we leveraged DOCSum during the development phase to assist in selecting hyperparameters.

.4.1.3 Dataset Overview

- **Initial Training Dataset** consists of 500 documents from PubMed and annotated with synthetic labels [Ref 10]. This dataset is used to train the base model (θ_0) before AL begins.
- **Active Learning Data Pool c**omprises 42 HDE-labeled (Gold Standard data) documents. These documents are selected based on uncertainty measures and compared with the HDEs annotation results before being added to the updated training dataset.
- **Test Set** consists of 49 HDE labeled documents, which are untouched throughout the AL cycles. This dataset was used to evaluate the model's performance, ensuring that improvements observed during AL are not due to data leakage but reflect the model's genuine generalization capabilities.

  The HDE-labeled documents were annotated by three medical domain experts, and the Kappa rates were between 0.73 and 0.97[Ref 10, 22, 23].

4.2 Metrics used in the experiments

We used Recall, F1-score, and weighted F measure: fbeta_score [Ref 30] to assess model performance. These metrics are tracked across successive AL cycles to measure the model improvement as it incorporates human feedback. Recall is selected as the primary metric of interest, which is motivated by the goal of minimizing the risk of missing any key phrases. F1-score provides a more holistic view of model performance, and it balances both precision and recall. Therefore, we prioritize recall but also consider F1-score. Weighted F measure is used to provide additional perspectives for readers. We used beta equals 5 and 10, i.e., recall is 5 times or 10 times more important than precision [Ref 30].

4.3 Hyperparameters and Experiment Settings

To ensure fair and robust comparisons across all AL strategies, we tuned the following **hyperparameters** systematically and compared the results. We varied one hyperparameter at a time, while holding all others constant, to assess its impact on model performance. The best-performing values were then combined to form the final set of hyperparameters.

- **Number of AL Cycles**: 4
- **Number of Documents Added per Cycle**: 5
- **Number of Copies per Document per Cycle**: 1
- **BiLM Retrained**: Yes, the Language Model, BiLM, is retrained after each AL cycle
- **Sequence Length (BiLM)**: 10
- **Learning Rate**: 0.001
- **Dropout Rate**: 0.3
- **Repetitions**: To ensure the stability and reliability of our results, we repeated each experiment 25 times and reported the average recall and F1 score along with their 95% confidence intervals (mean ± 1.96 × standard error).

The number of AL cycles is a key hyperparameter that influences model improvement over time. Typically, AL continues until a predefined stopping criterion is satisfied [Ref 24]. In our study, we initially experimented with eight cycles but found that most performance gains occurred during the early stages, with subsequent cycles yielding diminishing returns. Based on this observation, we chose to include four cycles of AL.

Additionally, informed by our previous experiments [Ref 10], we found that incorporating 1–2 HDE-annotated documents for every 100 synthetically labeled documents yields optimal results when HDE annotations are limited. Given these considerations and a total of 20 usable HDE-annotated documents for AL, we naturally settled on four cycles and added five documents per cycle. Out of the 42 HDE-labeled documents in our AL data pool, we allocated 20 for AL and reserved the remaining 22 to maintain variability in document selection across different strategies. This was a critical design choice—had all 42 documents been used for each strategy, the experiments would have become deterministic, limiting our ability to compare selection behavior. By limiting each strategy to 20 documents, we ensured diversity in selections, enabling fairer and more informative comparisons across AL strategies.

4.4 Results

4.4.1 Evaluation results by different active learning strategies

The results evaluate the impact of different strategies on the rate and magnitude of the recall and F1-score improvements. Tables 2 (Token Entropy) and 3 (MTP) present the average Recall and F1-score metrics for each AL cycle and strategy (Table 1). Margin results were included in the Appendix (Table A2).

Table 2: Comparison of Token Entropy-based AL strategies across four cycles (θ_0–θ_4).

Evaluation results for token entropy (arithmetic average ± 1.96 standard error; n = 25)								
Active learning cycle	**DOCAVG REV ORDER**		**DOCAVG ACTUAL ORDER**		**KPSUM REV ORDER**		**KPSUM ACTUAL ORDER**	
	RECALL	**F1**	**RECALL**	**F1**	**RECALL**	**F1**	**RECALL**	**F1**
θ_0	0.431 ± 0.0228	0.534 ± 0.0180	0.433 ± 0.0237	0.537 ± 0.0189	0.428 ± 0.0189	0.536 ± 0.0139	0.439 ± 0.0177	0.543 ± 0.0110
θ_1	0.532 ± 0.0098	0.574 ± 0.0052	0.527 ± 0.0128	**0.576 ± 0.0053**	0.518 ± 0.0162	0.572 ± 0.0069	0.530 ± 0.0060	0.576 ± 0.0028
θ_2	0.517 ± 0.0154	0.568 ± 0.0101	**0.538 ± 0.0152**	0.576 ± 0.0076	**0.539 ± 0.0138**	0.577 ± 0.0081	0.525 ± 0.0103	0.577 ± 0.0049
θ_3	**0.536 ± 0.0164**	**0.578 ± 0.0109**	0.529 ± 0.0137	0.570 ± 0.0076	0.535 ± 0.0107	**0.581 ± 0.0077**	**0.544 ± 0.0091**	**0.586 ± 0.0056**
θ_4	0.518 ± 0.0120	0.572 ± 0.0072	0.518 ± 0.0180	0.571 ± 0.0124	0.521 ± 0.0163	0.570 ± 0.0092	0.535 ± 0.0103	0.586 ± 0.0058

Table 3: Performance of Minimum Token Probability (MTP)-based AL strategies across five cycles (θ_0–θ_4).

Evaluation results for MTP (arithmetic average ± 1.96 standard error; n = 25)												
	DOCAVG REV ORDER		**DOCAVG ACTUAL ORDER**		**KPSUM REV ORDER**		**KPSUM ACTUAL ORDER**		**KPAVG REV ORDER**		**KPAVG ACTUAL ORDER**	
	RECALL	**F1**	**RECALL**	**F1**	**RECALL**	**F1**	**RECALL**	**F1**	**RECALL**	**F1**	**RECALL**	**F1**
θ_0	0.45 ± 0.0166	0.546 ± 0.0099	0.435 ± 0.015	0.542 ± 0.0094	0.475 ± 0.0121	0.560 ± 0.0073	0.438 ± 0.0177	0.543 ± 0.0110	0.423 ± 0.0198	0.532 ± 0.0144	0.424 ± 0.0209	0.530 ± 0.0159
θ_1	0.520 ± 0.0140	0.573 ± 0.0067	0.518 ± 0.0122	0.572 ± 0.0072	0.525 ± 0.0163	0.570 ± 0.0095	0.526 ± 0.0060	0.577 ± 0.0028	0.524 ± 0.0150	0.574 ± 0.0076	0.522 ± 0.0140	0.569 ± 0.0106
θ_2	**0.543 ± 0.0203**	0.569 ± 0.0132	**0.546 ± 0.0145**	**0.583 ± 0.0066**	**0.540 ± 0.0185**	0.575 ± 0.0078	**0.543 ± 0.0103**	**0.586 ± 0.0049**	**0.538 ± 0.0132**	**0.578 ± 0.0056**	**0.532 ± 0.0152**	**0.574 ± 0.0085**
θ_3	0.529 ± 0.0130	**0.579 ± 0.0057**	0.519 ± 0.0109	0.576 ± 0.0066	0.536 ± 0.0142	**0.576 ± 0.0054**	0.534 ± 0.0091	0.585 ± 0.0056	0.528 ± 0.0115	0.574 ± 0.0058	0.528 ± 0.0102	0.578 ± 0.0064
θ_4	0.533 ± 0.0117	0.576 ± 0.0051	0.525 ± 0.0116	0.569 ± 0.0059	0.534 ± 0.0130	0.574 ± 0.0069	0.532 ± 0.0103	0.576 ± 0.0058	0.529 ± 0.0148	0.573 ± 0.0077	0.501 ± 0.0174	0.565 ± 0.0103

Across all four Token Entropy strategies, DOCAVG and KPSUM present consistent improvements in both Recall and F1 over the AL cycles. The DOCAVG actual order and KPSUM actual order strategies show slightly higher final F1 scores (0.586) by θ_3, suggesting better overall generalization. KPSUM reverse order shows relatively lower F1 gains compared to the actual order strategies. Overall, the trends indicate that actual order sampling tends to be slightly more effective than reverse order, and both DOCAVG and KPSUM aggregation methods yield comparable performance improvements over time.

The results by all four Margin-based strategies show consistent improvement in recall and F1 scores from θ_0 to θ_1, with most gains occurring early in the Active Learning process (Table A2).

Across all strategies using Minimum Token Probability (MTP), performance improves steadily through θ_4, with early gains between θ_0 and θ_1. KPAVG (Actual Order) and KPSUM (Actual Order) achieve the highest F1 scores by θ_2–θ_3 (≈0.574 and ≈0.586, respectively). KPAVG (Reverse) also performs well in early rounds, showing recall gains from 0.423 to 0.529, indicating that uncertainty in non-keyphrase tokens may still offer training value. DOCAVG strategies show more moderate improvements, suggesting that averaging token uncertainty across all tokens may dilute signal quality.

The random sampling results serve as a baseline for evaluating the effectiveness of the AL strategies. Random sampling shows an initial improvement in both recall and F1 (beta = 1) from θ_0 to θ_2, with recall peaking at θ_2 (0.587) before declining (Table 4). This fluctuation suggests inconsistency in sampling quality, and the lack of targeted selection of documents likely limits model improvement beyond early rounds. F1 (Beta = 1) remains relatively stable after θ_1, indicating general knowledge gain but lower efficiency compared to uncertainty-based strategies.

Table 4: Performance of Recall and F-1 for the Random Sampling across four active learning cycles (θ_0–θ_4).

RANDOM (arithmetic average ± 1.96 standard error; n = 25)				
	RECALL AVG	**F1 AVG (Beta = 1)**	**F1 AVG (Beta = 5)**	**F1 AVG (Beta = 10)**
θ_0	0.440 ± 0.0135	0.543 ± 0.0098	0.446 ± 0.0133	0.441± 0.0134
θ_1	0.523 ± 0.0135	**0.579 ± 0.0057**	0.527 ± 0.0129	0.524 ± 0.0134
θ_2	**0.587 ± 0.0108**	0.577 ± 0.0047	**0.586 ± 0.0099**	**0.587 ± 0.0106**
θ_3	0.519 ± 0.0135	0.571 ± 0.0088	0.523 ± 0.0127	0.520± 0.0130
θ_4	0.509 ± 0.0172	0.574 ± 0.0112	0.514 ± 0.0168	0.510 ± 0.0171

4.4.2 Evaluation results comparison across different active learning strategies

Figure 3a presents the Recall, and Figure 3b presents the F1 progress across AL cycles for Token Entropy. We noticed that most of the performance improvement occurs between the initial model (θ_0) and the first cycle (θ_1). Towards the later cycles, the results exhibit fluctuations in recall, suggesting that performance gains begin to plateau. F1 scores show similar trends to the recall. Our results demonstrate a slightly larger improvement for recall than for F1 scores between θ_1 and θ_0. Figure 3c and 3d illustrate the average Recall (3c) and F1 Scores (3d) across AL cycles for the Token Entropy KPSum Actual strategy compared to Random sampling. Figure 3c and 3d highlight the tendency of Random selection to produce early but less stable gains, while Active Learning via KPSum Actual leads to more gradual but sustained improvements in both Recall and F1 scores. The figures for comparing Recall and F1 scores for strategies of margin and MTP show similar results to Token Entropy and are included in the Appendix (Figures A1 to A4).

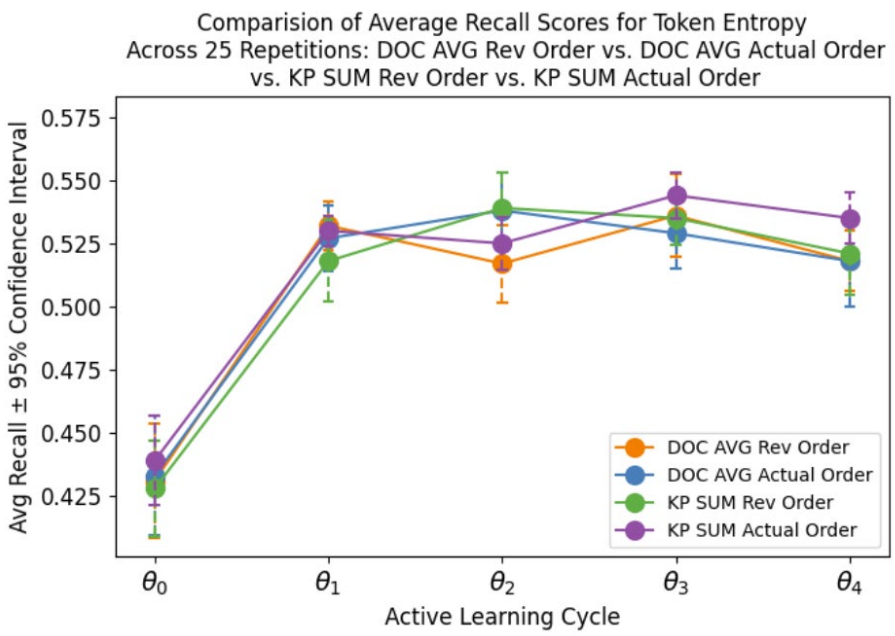


Figure 3(a): Recall progress across Active Learning cycle for Token Entropy

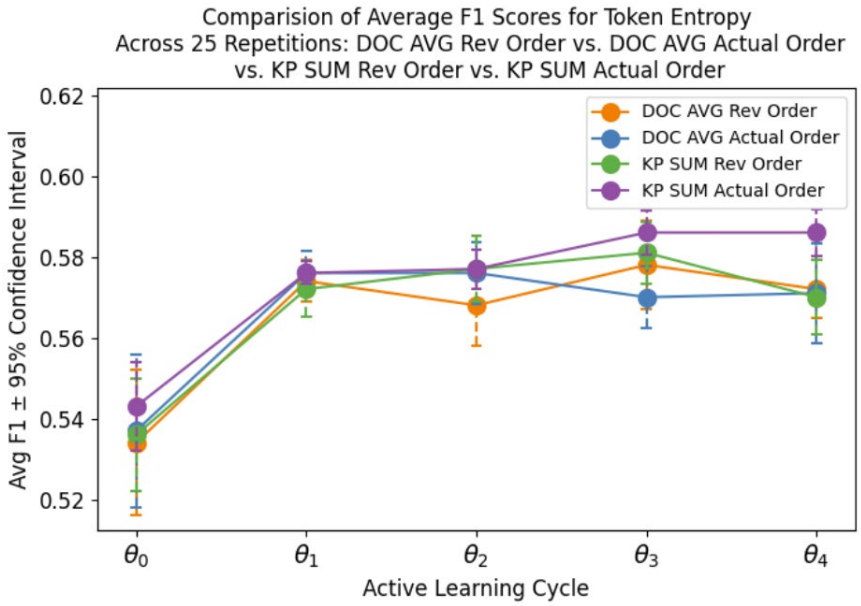


Figure 3(b): F1 score progression across Active Learning cycles for Token Entropy

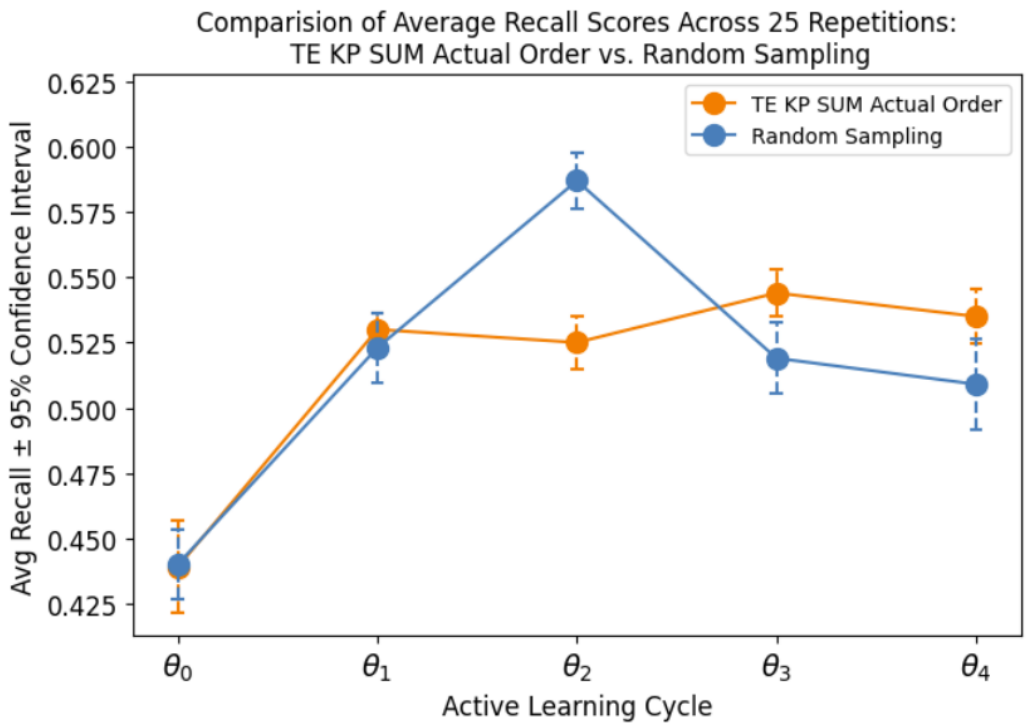


Figure 3(c): Average Recall comparison of TE KpSum Actual and Random Sampling

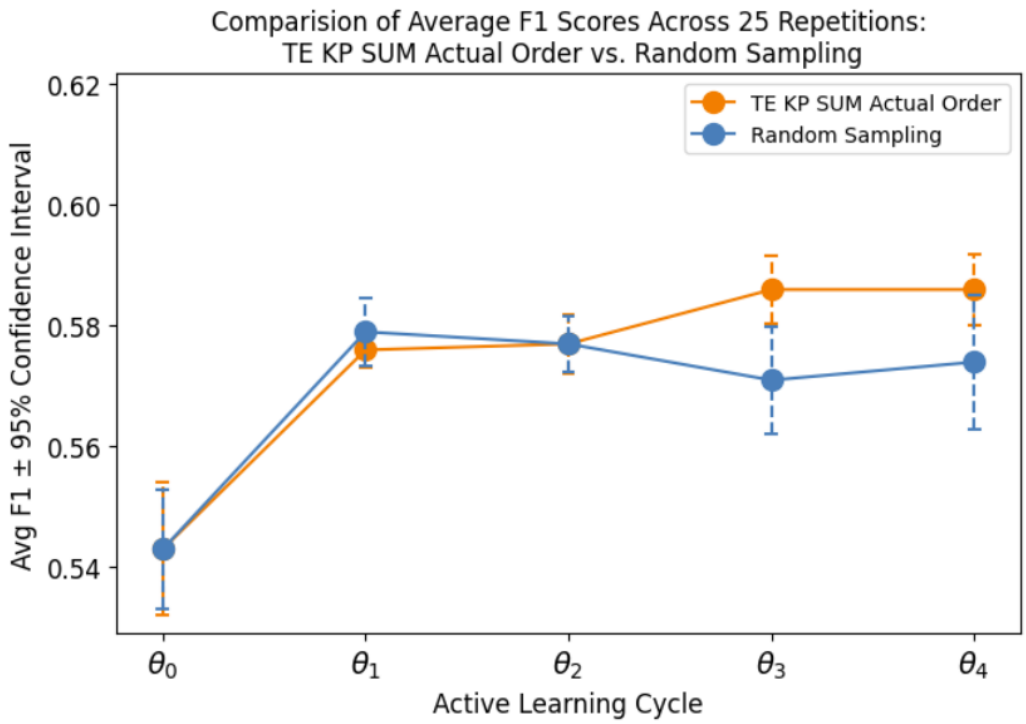


Figure 3(d): Average F1 comparison of TE Kpsum Actual and Random Sampling

4.4.3 Original and weighted F1 values by different active learning strategies

Table 5 reports the weighted F1 scores (± 1.96 standard error) for Token Entropy–based AL strategies evaluated at β = 1, β = 5, and β = 10 across five AL cycles (θ_0 to θ_4). At θ_0, all strategies start with comparable scores, ranging from 0.430 to 0.445. DOCAVG Actual Order shows consistent performance, reaching 0.542 (β = 5) at θ_2. KPSUM Actual Order attains the highest score of 0.546 (β = 5) at θ_2, while also maintaining stable results across other cycles. KPSUM Reverse Order peaks at θ_2 with 0.542 (β = 5) and slightly declines in later cycles. DOCAVG Reverse Order sees a rise from 0.439 (θ_0, β = 5) to 0.541 (θ_2, β = 5), then slightly fluctuates. Across all strategies and cycles, the β = 5 and β = 10 values remain closely aligned, differing only marginally within each configuration. The weighted F1 scores for the Random sampling are in Table 4. The results suggest that while Random sampling provides early performance gains, it lacks consistency in maintaining or improving performance in later stages of the Active Learning process. The weighted F1 scores for margin and MTP are included in the Appendix (Tables A3 and A4).

Table 5: Weighted F1 scores for Token Entropy–based Active Learning strategies, evaluated at β = 1, β = 5, and β = 10.

COMBINED WEIGHTED F1 FOR TOKEN ENTROPY (arithmetic average ± 1.96 standard error; n = 25)												
	DOCAVG REV ORDER			**DOCAVG ACTUAL ORDER**			**KPSUM REVERSE ORDER**			**KPSUM ACTUAL ORDER**		
	Beta = 1	**Beta = 5**	**Beta = 10**	**Beta = 1**	**Beta = 5**	**Beta = 10**	**Beta = 1**	**Beta = 5**	**Beta = 10**	**Beta = 1**	**Beta = 5**	**Beta = 10**
θ_0	0.534 ± 0.0180	0.439 ± 0.0227	0.434 ± 0.0228	0.537 ± 0.0189	0.440 ± 0.0236	0.435 ± 0.0237	0.536 ± 0.0139	0.434 ± 0.0186	0.430 ± 0.0188	0.543 ± 0.0110	0.445 ± 0.0174	0.440 ± 0.0177
θ_1	0.574 ± 0.0052	0.523 ± 0.0094	0.52 ± 0.0097	**0.576 ± 0.0053**	0.529 ± 0.0121	0.526 ± 0.0126	0.572 ± 0.0069	0.522 ± 0.0156	0.519 ± 0.0160	0.576 ± 0.0028	0.523 ± 0.0058	0.527 ± 0.0060
θ_2	0.568 ± 0.0101	**0.541 ± 0.0150**	**0.539 ± 0.0153**	0.576 ± 0.0076	**0.542 ± 0.0146**	0.526 ± 0.0150	0.577 ± 0.0081	**0.542 ± 0.0134**	**0.540 ± 0.0137**	0.577 ± 0.0049	**0.546 ± 0.0099**	**0.544 ± 0.0102**
θ_3	**0.578 ± 0.0109**	0.523 ± 0.0160	0.521 ± 0.0163	0.570 ± 0.0076	0.531 ± 0.0132	**0.529 ± 0.0136**	**0.581 ± 0.0077**	0.539 ± 0.0104	0.536 ± 0.0106	**0.586 ± 0.0056**	0.538 ± 0.0088	0.535 ± 0.0090
θ_4	0.572 ± 0.0072	0.535 ± 0.0116	0.533 ± 0.0119	0.571 ± 0.0124	0.520 ± 0.0177	0.517 ± 0.0179	0.570 ± 0.0092	0.525 ± 0.0157	0.522 ± 0.0162	0.586 ± 0.0058	0.535 ± 0.0099	0.533 ± 0.0102

5 Discussion

After examining the full set of AL results, we share the following observations with peers and hope they can be reference points in other AL work.

5.1 Majority of Performance Gains Occur from θ_0 to θ_1.

We noticed that most of the performance improvement occurs during the transition from θ_0 to θ_1, regardless of the strategies. This pattern aligns with findings from previous studies [Ref 25]. This highlights the importance of the initial rounds of feedback, as they address the most critical gaps in the model's learning, leading to rapid improvements. It emphasizes the need to prioritize high-quality labels and diverse data during the initial cycles to maximize the impact of the active learning process. When documents are added randomly, there is still a significant improvement from θ_0 to θ_1 in both F1 and recall metrics. This indicates that while Active Learning strategies optimize sample selection, the introduction of HDE-labeled data itself is a major factor in the observed performance boost. Another unique characteristic of our use case is the complexity of the task, which is significantly different from a typical binary predictive task.

5.2 Identifying Saturation in Active Learning Cycles

Determining an effective stopping criterion is crucial. A well-defined stopping condition prevents excessive labeling costs and mitigates the risk of overfitting the AL model. One approach is to monitor the convergence of learning by evaluating the generalization error on a test dataset, as discussed in H Ishibashi, et al *[Ref 24]*. However, obtaining a sufficiently large and representative test set may not always be feasible. A practical alternative is to define saturation-based stopping criteria, where the performance improvements fall below a predefined threshold (e.g., less than 1% improvement in recall or F1-score) or when key evaluation metrics begin to fluctuate around a stable value.

As shown in most of the strategies presented in Tables 2 and 3 and Figures 3, the recall generally increases from θ_0 to θ_2, after which it either plateaus or begins to fluctuate slightly without consistent gains. This trend suggests that the model makes its most substantial improvements during the early cycles of active learning. A similar trend is observed in F1 scores. However, it is important to note that we may be observing saturation earlier in our experiments due to the limited size of our AL data pool.

5.3 Impact of KP SUM Strategies on Recall and F1 Improvement in Active Learning

In Tables 2, 3, and A2 the most consistent and stable improvements in recall are observed for KPSUM Actual, with the highest recall gains and lower variation in the confidence intervals, indicating more stable learning. No other strategies exhibit the same level of consistency across all metrics and sampling techniques. The effectiveness of KPSUM lies in its construction: it calculates document uncertainty solely based on key phrases, assigning higher scores to documents with a greater density of such phrases. The model is repeatedly exposed to examples that reinforce the structure, context, and patterns associated with key phrases. This focused exposure leads to more robust generalization and improved recall performance. Although the per-cycle recall gain of KPSUM Actual may not always be dramatically higher than other strategies, its sustained improvement and model stability suggest a long-term advantage.

5.4 F1 and Recall Dynamics in KP AVG Strategies: Actual vs. Reverse

KP AVG Reverse consistently achieves higher F1 scores than KP AVG Actual across cycles, as shown in Tables 2, 3 and A2 and Figure A5 for MTP uncertainty measure. This suggests that selecting documents with lower keyphrase uncertainty (as done in Reverse) helps the model learn to reduce false positives, improving precision and overall F1. Although KP AVG Actual was expected to improve recall by selecting documents with higher keyphrase uncertainty, the observed recall improvements are marginal and inconsistent compared to the Reverse strategy. This indicates that KP AVG Actual in exposing the model to uncertain keyphrase boundaries may not be strong enough to translate into significantly better recall.

5.5 Comparing Early-Stage Improvements: Reverse vs. Actual Order

We hypothesized that Reverse order would yield quicker improvements in the early stages (θ_0 to θ_1) and that Actual order would show more sustained improvements later (θ_0 to θ_2 or θ_3). This

is why we experimented with both. However, the observed pattern did not hold across all strategies but showed advantages in some strategies, possibly due to the limited dataset. While the Reverse order shows quicker gains initially, the Actual order did not consistently lead to stronger long-term improvements as expected. With a larger dataset, the actual order's advantage may become more prominent.

5.6 Active Learning vs. Random Selection: Impact on Recall and Model Stability

Our experiments show that while both the AL (KPSUM Actual) and Random selection strategies yield improvements in performance, the difference lies in the stability of these improvements. KPSUM Actual demonstrates more reliable gains, especially after the initial learning cycles. The F1 and Recall scores for KPSUM Actual show a more consistent upward trend compared to Random selection. In contrast, Random selection shows improvements initially, but exhibits volatility. After a certain point, the recall and F1 scores for Random selection can decline, highlighting the inefficiency of random document addition in refining the model's performance. These results align with findings from 2021 [Ref 25]. Our results further underscore the value of a structured, uncertainty-based approach for document selection, such as KPSUM Actual, which offers more reliable and stable performance gains than Random selection.

5.7 F1 scores and weighted F1 scores

F measures are typically used in machine learning, information retrieval, and NLP tasks. Although in our use case, recall is much more important than F measure, we still keep F measure in both original values and weighted values to provide a more balanced and comprehensive measure for model performance. From Table 5, A3 and A4, we noticed weighted F1 scores (beta = 5, 10) show a similar trend. Weighted F1 values have shown approximately 0.03 to 0.05 lower than the original F1 values across different strategies. Moreover, weighted F1 values peaked at θ_2 constantly across most strategies, including random sampling. The original F1 value peaked inconsistently, including θ_1, θ_2, and θ_3. Overall, weighted F1 values provide important complementary measures to original F1 values and recalls.

5.8 How will the results be used?

We started the investigation aiming to facilitate the long-term maintenance of the CDSS ontology. The results presented in this paper show promise; however, they may not be ideal yet. Considering the complexity of our task: 1) not a binary prediction, 2) the three HDEs could not reach 100% consensus even after discussions, we are very delighted with the initial exploration and the results. Realistically, workflow automation is necessary to make the long-term ontology maintenance feasible; however, considering the quality requirements of ontology, the current technology has not reached the level that can be relied on solely in operation. This can also be demonstrated by the work using LLMs to extract semantics [Ref 15], which demonstrated the exploration of using the most advanced technology to achieve similar goals, however, with comparable results to ours. This indicates the task is indeed challenging and there is plenty of space for improvement. Therefore, putting HDE in the loop for our task is necessary in the

foreseeable future. We think that we are on the right track to use active learning to facilitate the task. Meanwhile, continuing to explore ways to improve the model performance is a needed future priority.

5.9 Limitations and Future Work

Our study has several limitations: 1) the performance of the current active learning pipeline can be improved further. The model's performance needs improvement to be useful in real-world operation. 2) In this study, we have a limited number of ground truth documents. In the future, with a larger number of ground truth documents, others may obtain different results. However, we believe the approach described in this paper can still be used. 3) In this study, we explored basic aggregation strategies for estimating document-level uncertainty, primarily focusing on sum-based (e.g., KPSum, DocSum) and average-based (e.g., KPAvg, DocAvg) methods. The future work can include advanced mathematical compositions, such as exponential or logarithmic calculations. In addition, future work could explore alternative aggregation techniques such as trimmed mean or weighted mean. Techniques like median, mode, or box-plot-based selection (e.g., interquartile range filtering) may also provide more robust document representations, especially in skewed distributions. In addition, the real impact of active learning in HDE annotation tasks in the ontology construction and maintenance context can also be measured based on the work saved over sampling algorithms [Ref 26, 27, 28, 29].

In addition, although prior studies such as [31] and [32] have examined token-level uncertainty in NER and uncertainty-based instance selection more broadly, our work extends these efforts by proposing task-specific document-level aggregation techniques for Active Learning in a keyphrase identification context. This represents a novel abstraction layer between token uncertainty and document selection—especially important in ontology construction and maintenance, where annotation decisions occur at the document level.

Section 6: CONCLUSION

In this study, we proposed and evaluated new measures for an Active Learning framework to identify candidate keyphrases for a CDSS ontology using a BiLSTM-CRF model, by incorporating HDE feedback. A key contribution of our work is the novel document-level uncertainty aggregation strategies (KPSum, KPAvg, DocSum, and DocAvg), which enable more effective, transparent, and reproducible document selection for annotation. Based on our experimental results, KPSum (Actual) emerged as the most effective strategy for improving keyphrase identification performance in our setting. Meanwhile, the other aggregation methods presented in this manuscript provide a foundation for peers to compare and select the optimal approach and measures for their tasks and datasets. Through systematic comparison of various Active Learning strategies and uncertainty measures, this study offers practical benchmarks for future research in similar domains. While our experiments centered on keyphrase identification, the proposed framework is domain-agnostic and can be adapted to other sequence labeling tasks. Overall, this work contributes to the underexplored application of Active Learning in deep learning-based keyphrase identification, providing a scalable and adaptive foundation for efficient HDE annotation in NLP.

Acknowledgement

This study is made possible by funding from the National Institute of General Medical Sciences (R01GM138589) at the National Institutes of Health and partially through P20GM121342 of NIGMS and partially through # OIA-2242812 of the National Science Foundation.

Data Availability

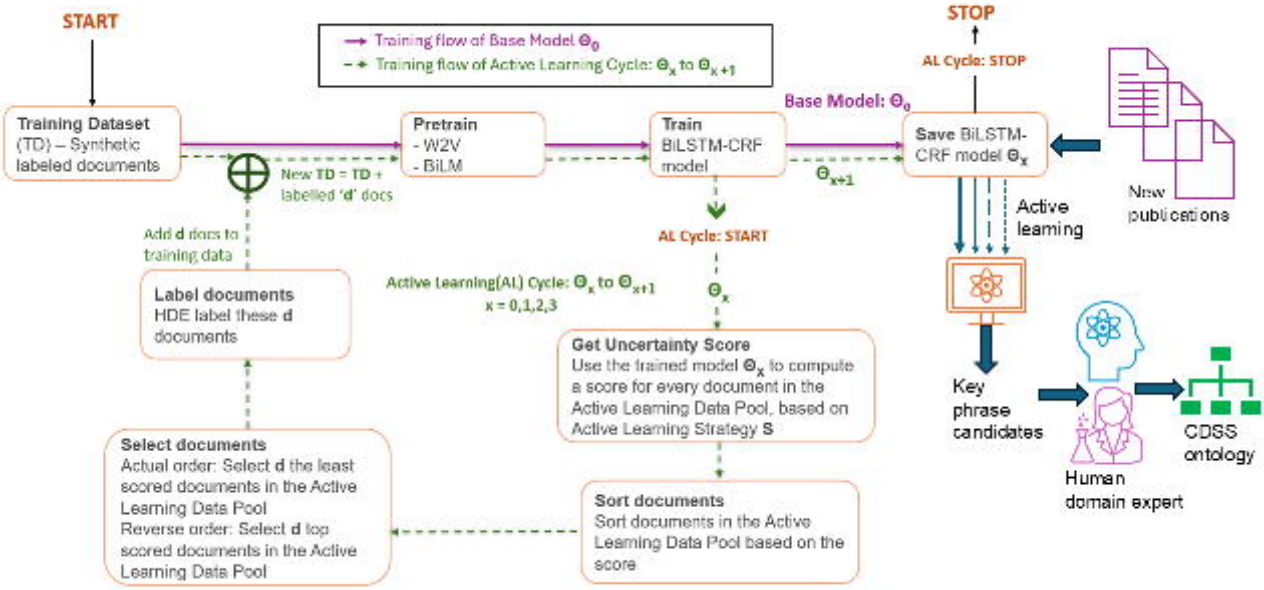
The full set of evaluation results and a full version of the manuscript can be found in the link: https://github.com/CDSS4PCP/NLP_KPIdentify/tree/active_learning_files/ActiveLearningStrategiesResults

REFERENCES:

1. Greenes, R. and G.D. Fiol, Clinical Decision Support and Beyond: Progress and Opportunities in Knowledge-Enhanced Health and Healthcare. 3rd Edition ed. 2023, Cambridge, MA: Elsevier: Academic Press.
2. Davies, J., D. Fensel, and F.V. Harmelen, Towards the semantic web: Ontology-driven knowledge management. 2003, Chichester: John Wiley & Sons.
3. Supekar, K., et al., Knowledge zone: A public respiratory of peer-reviewed biomedical ontologies in MEDINFO 2007, K.e.a. Kuhn, Editor. 2007, IOS Press: Brisbane, Australia. p. 812-816. Rector, A.L. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. in K-CAP'03. 2003. Sanibel Island, Florida, USA.
4. Rector AL. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. Proceedings of the 2nd international conference on Knowledge capture. 121-128. K-CAP'03; 2003 October 23–25, 2003; Sanibel Island, Florida, USA
5. Liu, M., Tu, Z., Zhang, T. *et al.* LTP: A New Active Learning Strategy for CRF-Based Named Entity Recognition. *Neural Process Lett* 54, 2433–2454 (2022). https://doi.org/10.1007/s11063-021-10737-x
6. Rose, S.J., Engel, D.W., Cramer, N., & Cowley, W. (2010). 1 Automatic keyword extraction from individual documents. (TF IDF for Keyword retrieval)
7. Lafferty, J.D., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *International Conference on Machine Learning*. (CRF for sequence tagging)
8. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv, abs/1508.01991*.(BiLSTM CRF Model)
9. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.(BERT)
10. Goli R, Hubig N, Min H, Gong Y, Sittig D, Robinson D, et al. Keyphrase Identification Using Minimal Labeled Data with Hierarchical Context and Transfer Learning. BMC Medical Informatics and Decision Making. Under review. MedRxiv Preprint. 2023. https://medrxiv.org/cgi/content/short/2023.01.26.23285060v2

11. Monarch, R.M., Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. 2021, Shelter Island, NY: Manning Publications.
12. Settles, B. (2009). Active Learning Literature Survey.
13. Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2630. URL https://www.aclweb.org/anthology/ W17-2630.
14. 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.
15. Caufield J, Hegde H, Emonet V, Harris N, Joachimiak M, Matentzoglu N. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. Bioinformatics. 2024;40(3).
16. 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..
17. 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.
18. S. -J. Huang, R. Jin and Z. -H. Zhou, "Active Learning by Querying Informative and Representative Examples," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 1936-1949, Oct. 2014, doi: 10.1109/TPAMI.2014.2307881.keywords: {Measurement uncertainty;Uncertainty;Algorithm design and analysis;Correlation;Labeling;Clustering algorithms;Kernel;Active learning;learning with unlabeled data;multi-label learning;informativeness;representativeness}.
19. Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. 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.
20. Dara Bahri, Heinrich Jiang, Tal Schuster, and Afshin Rostamizadeh. 2022. Is margin all you need? An extensive empirical study of active learning on tabular data, arxiv. DOI: https://arxiv.org/abs/2210.03822
21. Neha Gupta, Hari Narasimhan, Ankit Singh Rawat, Wittawat Jitkrittum, Aditya Menon, and Sanjiv Kumar. 2024. Language Model Cascades: Token-Level Uncertainty And Beyond, arxiv. DOI: https://doi.org/10.48550/arXiv.2404.10136

22. Jing X, Goli R, Komatineni K, Alluri S, Hubig N, Min H, et al. Active Learning Pipeline to Identify Candidate Terms for a CDSS Ontology. Stud Health Technol Inform. 2024;316:1338-42. DOI: 10.3233/shti240660
23. Jing X, Indani A, Hubig NC, Min H, Gong Y, Cimino JJ, et al. A systematic approach to configuring MetaMap for optimal performance. Methods Inf Med. 2022. DOI: 10.1055/a-1862-0421
24. Ishibashi, H., & Hino, H. (2020). Stopping criterion for active learning based on deterministic generalization bounds. *ArXiv, abs/2005.07402*.
25. Mosqueira-Rey, E., Alonso-Ríos, D. & Baamonde-Lozano, A. Integrating Iterative Machine Teaching and Active Learning into the Machine Learning Loop. *Procedia Comput. Sci.* 192, 553–562 (2021) .
26. Wallace BC, Small K, Brodley CE, Trikalinos TA. Active learning for biomedical citation screening. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; Washington, DC, USA: Association for Computing Machinery; 2010. p. 173–82.
27. Wallace BC, Small K, Brodley CE, Trikalinos TA. Who Should Label What? Instance Allocation in Multiple Expert Active Learning. Proceedings of the 2011 SIAM International Conference on Data Mining (SDM). p. 176-87.
28. Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, Trikalinos TA, editors. Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews. Proceedings of the 1st ACM International Health Informatics Symposium; 2010 Nov 11-12, 2010; Arlington, VA.
29. Byron C. W, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11:55-67.
30. Scikit-Learn. fbeta_score 2025 https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html.
31. Hoarau, Arthur & Lemaire, Vincent & Gall, Yolande & Dubois, Jean-Christophe & Martin, Arnaud. (2024). Evidential uncertainty sampling strategies for active learning. Machine Learning. 113. 1-22. 10.1007/s10994-024-06567-2.
32. Nguyen, Vu-Linh & Shaker, Mohammad & Hüllermeier, Eyke. (2022). How to measure uncertainty in uncertainty sampling for active learning. Machine Learning. 111. 10.1007/s10994-021-06003-9.


START
Training flow of Base Model Θ_0
Training flow of Active Learning Cycle: Θ_x to Θ_{x+1}
STOP
AL Cycle: STOP
Base Model: Θ_0
Training Dataset
(TD) – Synthetic labeled documents
Pretrain
- W2V
- BiLM
Train
BiLSTM-CRF model
Save BiLSTM-CRF model Θ_x
Θ_{x+1}
New TD = TD + labelled 'd' docs
Active learning
New publications
Add d docs to training data
AL Cycle: START
Active Learning(AL) Cycle: Θ_x to Θ_{x+1}
x = 0,1,2,3
Θ_x
Label documents
HDE label these d documents
Get Uncertainty Score
Use the trained model Θ_x to compute a score for every document in the Active Learning Data Pool, based on Active Learning Strategy S
Key phrase candidates
Human domain expert
CDSS ontology
Select documents
Actual order: Select d the least scored documents in the Active Learning Data Pool
Reverse order: Select d top scored documents in the Active Learning Data Pool
Sort documents
Sort documents in the Active Learning Data Pool based on the score