Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-021-00230-w

REGULAR PAPER

Active learning approach using a modified least confidence sampling strategy for named entity recognition

Ankit Agrawal[1] · **Sarsij Tripathi**[2] · **Manu Vardhan**[1]

Received: 19 December 2019 / Accepted: 5 January 2021

Abstract
One of the important subtasks of information extraction is named entity recognition (NER). Its aim is to identify and to classify the named entities in the textual data into predetermined categories. There are a large number of supervised learning and deep learning models being developed for the entity recognition task, which performs well in the presence of a labeled training set. The availability of the labeled training set requires the labeling of large unlabeled data, which is both expensive and time taking. Active learning is an iterative approach that provides a way to minimize labeling cost without affecting performance. This approach uses a sampling strategy that selects the appropriate unlabeled data instances, an oracle to label the selected data instances, and a machine learning model (base classifier). In this work, a modified least confidence-based query sampling strategy for the active learning approach for named entity recognition task has been proposed, which considers different numbers of uncertain words present within the sentences to compute the final least confidence score of the sentence for comparison. To evaluate the effectiveness of the proposed approach, the comparison of the performance is made among the active learning approaches with the proposed sampling strategy, random sampling strategy, and two other well-known existing uncertainty query sampling strategies. Real-world scenario for active learning approach is simulated for experiment, and the total amount of labeled data required for training of active learner to reach the stop condition while using different sampling strategies is recorded. The experiment is carried for the development and the test set of the three different biomedical corpora and a Spanish language NER corpus. It is found that with the proposed active learning approach, there is a minimal requirement of labeled data for training to reach the above performance level in comparison with the other approaches. The performance of the proposed approach is found to be slightly better than the existing sampling approach, and the performance of all the approaches is far better than the random sampling approach.

Keywords Named entity recognition · Active learning · Least confidence · Sampling strategy · Supervised learning

1 Introduction

There is exponential growth in textual data over the Internet from various social media platforms in various forms such as scientific journals, e-books, web pages, news, learning content, e-mail, etc. Also, there is tremendous growth in the number of scientific research literature with enormous new publications in millions each year. There is information overload for biomedical researchers due to the exponential growth of biomedical data. At the time of writing this research article, there are 5.7 million biomedical and life sciences literatures archived in PMC alone [1]. All these textual data are present in digital text form, and often these data are required to be analyzed so that important information can be extracted. There are a number of natural language processing (NLP) applications such as question answering

✉ Ankit Agrawal
aagrawal.phd2017.cse@nitrr.ac.in

Sarsij Tripathi
sarsij@mnnit.ac.in

Manu Vardhan
mvardhan.cs@nitrr.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology Raipur, Raipur, Chhattisgarh, India

[2] Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, Uttar Pradesh, India

Published online: 19 January 2021

[2, 3], paraphrase detection [4], information retrieval [5], information extraction [6], machine translation [7], text summarization [8], etc., which make the use of named entity recognition as a subtask. Named Entity Recognition (NER) is an important subtask of the information extraction which deals with detection and identification of named entities present in text. According to [9], named entity recognition (NER) performs two tasks, i.e., locating the named entities within the text and tagging them with predefined classes, such as organization, location, person name, etc. Some of the popular machine learning algorithms used commonly in the past for NER include conditional random field (CRFs) [10], hidden markov model (HMM) [11], support vector machine (SVM) [12], maximum entropy (ME) [11], neural network-based models [13], etc. [14]. These models typically perform well when trained over large amount of labeled data and have shown improvement in NER for various domains, including the biomedical domain. In this paper, the conditional random field model has been used, which has been used commonly by researchers for the NER task. However, the proposed approach will also work with any of the above supervised machine learning models.

For biomedical research, information extraction has become an important tool due to exponential growth in the number of biomedical literature. The Biomedical Named Entity Recognition (BNER) is treated as a problem of sequence labeling, for which the aim of the machine learning models is to identify important biomedical entities present in the text such as species, diseases, genes, and gene products. Correct identification of biomedical concepts (or Biomedical Named Entity Recognition) is the basis for the development of biomedical NLP systems. The biomedical text possesses certain characteristics which makes the task of entity recognition challenging, such as abbreviations that are ambiguous (e.g., 'TCF' may refer to 'Tissue Culture Fluid' or to 'T cell factor'), multiple spelling forms (e.g., 'NAcetylCysteine', 'N-acetyl-cysteine' and 'N-acetylcysteine'), common head nouns (e.g., '84 and 91 kDa proteins' refers to '84 kDa protein' and '91 kDa protein') [15]. A large number of approaches have been proposed for the Named Entity Recognition (NER) task in the biomedical field and in other language specific domains, mostly for which a large amount of annotated data is available. Most of these approaches use linguistic, orthographic, contextual, and external evidence to extract the different patterns and features present in the data such that it indicates the named entities (NE). However, there is a need of large amount of annotated data for the high-performing NER approaches. The annotation of the large un-annotated dataset is very costly and time-consuming. There is a requirement of subject experts to perform the annotation, and it requires a lot of time and effort.

In supervised learning approach, the machine learning model is simply trained over the available labeled training set. The active learning approach provides a solution for the problem of nonavailability of labeled data up to an extent, and it is an iterative approach in which the active learner selects the most informative instances of unlabeled data instances for training using a query strategy. The selected informative instances of unlabeled data are the data instances from which active learner can learn the most. The model is trained iteratively using the active learning approach and tries to achieve higher accuracy with significantly less amount of labeled data required for training in comparison with that of supervised learning [16].

In this paper, we have proposed an active learning approach for Named Entity Recognition task using modified least confidence query strategy that considers the different number of most uncertain words present in unlabeled sentences of the dataset for calculation of final uncertainty scores, which is further used for selection of most informative unlabeled data instances. The experiments have been carried over three different biomedical datasets of different areas and a Spanish language dataset. We have compared our proposed approach with the marginal query sampling strategy, entropy query sampling strategy, and the baseline random query sampling strategy for active learning on the basis of the requirement of labeled data for training of active learner to reach to certain specific performance level. The sample code of proposed approach is available at https://github.com/ankitphd/modified-least-confidence-sampling-strategy-for-ner.

2 Related works

In this section, the existing research work related to the active learning approach for named entity recognition task has been discussed. The active learning experiment is simulated by [17] that uses the existing and new query sampling strategies belonging to three different categories, i.e., baseline sampling strategies, uncertainty-based strategies, and diversity-based strategies. The performance of the above query strategies was compared with the random sampling strategy in terms of annotation cost. They found that uncertainty-based strategies performed best in comparison with other strategies. Hence, in this research work, we have also compared the proposed approach with two other well-known uncertainty-based query sampling strategies.

In another work by [18], two different machine learning models have been used with a marginal query sampling approach. They performed an experiment for the active learning approach with machine learning models, which includes SVM, CRF, and the ensemble of both SVM and CRF. The experiment has been carried for three different language corpora and a biomedical corpus. They have tested their approach after training over a complete labeled train set

by selecting sentences from an unlabeled development set, that too for only ten iterations.

In a recent work, a new active learning approach is proposed by [19], which uses BERT-CRF model for named entity recognition task. They have used three different query sampling strategies, which include lowest token probability, least confidence, and normalized least confidence. In this paper, the authors have discussed the shortcoming of the least confidence strategy, i.e., least confidence strategy favors longer sentences (as summation takes place over tokens). Hence, in this research work, a modified least confidence query sampling approach is proposed, which is able to consider the different number of tokens within the sentences of the unlabeled text dataset for calculation of the final uncertainty score.

In general, a neural network-based model requires a large amount of labeled data to perform well. However, [13] has proposed a new deep active learning approach in which they have used a lightweight architecture based on CNN-CNN-LSTM model for named entity recognition. The model achieved state-of-the-art results for popular named entity recognition datasets while using only about 25% of the labeled train data.

A new active learning-based approach is proposed by [20] for developing a low-cost named entity recognition model. In the proposed approach, the sentences are first clustered using the k-means clustering algorithm. Further, the most suitable unlabeled data instances are chosen from the clustered unlabeled sentences using stratified sampling and entropy-based query strategy. The proposed approach is applied to extract legal and medical entities where it is found that the approach is capable to speed up the process and reduce the annotation cost significantly.

A package is also introduced by [21], which provides a library to ease the annotation task by minimizing the need for annotated data for the training of machine learning classifiers. It makes use of pre-annotation and active learning to achieve this task and can adapt to various annotation formats. A similar platform named INCEpTION is used for web-based annotation [22]. This platform can be used for various tasks, including entity linking, populating knowledge base, and semantic frame annotation [23]. The platform combines the annotation capabilities with the machine learning approaches, which in turn is useful for enhancing annotation efficiency. It has an active learning mode in which the learner uses the confidence scores assigned by the recommender.

Another work for clinical named entity recognition is presented by [24]. Multiple active learning query strategies have been investigated in depth, and their performance has been compared with random sampling baseline and supervised learning. They found that the active learning approach can reduce 35% annotation time and 20% further when started from scratch. Similar other work related to active learning for named entity recognition has been done for the social media text [25, 26]. In most of the research work carried out for the active learning approach for named entity recognition task, the researchers used the CRF model. The neural network models are iterative themselves, and they also need a large amount of labeled train data to perform well. Hence, in this paper, we have also used the conditional random field model as a baseline active learner for all the different active learning query sampling strategies.

3 Corpus details

Three different corpora are used for this experiment from the biomedical domain, which are popular and used earlier in research papers. The corpora are JNLPBA corpus, NCBI disease corpus, and BioCreative V Chemical Disease Relation (BC5CDR) task corpus present in CoNLL IOB2 format (i.e., Inside-Outside-Begin of the tag). All of these corpora belong to different biomedical domains. The first corpus is the JNLPBA corpus, used in BioNLP/JNLPBA Shared Task 2004 [27, 28], which was organized by the GENIA project. It is one of the most widely used corpora for entity recognition task in the field of molecular biology. The JNLPBA corpus is annotated in five different categories: cell type, cell line, RNA, DNA, and protein. This corpus is a simplified version of the GENIA version 3.02 corpus [29], originally annotated in 36 different categories. The JNLPBA corpus consists of 2000 MEDLINE abstracts in the train set, and 404 randomly selected MEDLINE abstracts in the test set. For this experiment, we have a development set used by [30] that contains 200 abstracts out of 2000 abstracts in the train set. The details of the JNLPBA corpus used in our active learning approach are presented in Table 1.

Table 1 Details of JNLPBA corpus used in experiment

Characteristics	Train set	Development set	Test set	Total classes (excluding Others tag)
No. of abstracts	1800	200	404	Five classes (Protein, DNA, RNA, cell-type and cell-line)
No. of sentences	18,607	1939	4260	
No. of tokens	446,890	47,661	101,443	

The second corpus is the NCBI disease corpus, which contains 793 PubMed abstracts and has a total of 6892 disease mentions that fall into 790 unique concepts [30, 31]. The details of the NCBI disease corpus are as follows in Table 2.

The third corpus used in this experiment is the BioCreative V Chemical Disease Relation (BC5CDR) task corpus. The corpus consists of 1500 PubMed articles manually labeled by experts for diseases, chemical, and their relationship [30, 32]. The corpus used for this experiment consists of the annotation in chemical and disease categories. Further details related to the BC5CDR corpus are presented in Table 3. The details of the number of abstracts for each of the three datasets are taken from [30, 33].

The fourth corpus used in this experiment is carried over the widely popular Spanish CoNLL 2002 NER dataset. The Spanish data present in the dataset are provided by the Spanish EFE News Agency which comprise Spanish newswire articles [34]. Other details of the Spanish CoNLL 2002 corpus are presented in Table 4.

The sample from the NCBI dataset, i.e., a sentence and respective labels, can be seen in Table 5. Only the sentences and the labels are present in the original datasets.

4 Methodology

In this section, various features used for the experiment are discussed in detail. Also, the baseline classifier and the evaluation metrics used for the named entity recognition task are discussed.

4.1 Feature set

The performance of any machine learning model depends mostly upon the feature set used for the task. It makes it very important to extract and pass the appropriate features with respect to the task and machine learning model. For our active learning-based approach for named entity recognition task, we have extracted the features commonly used by the researchers for named entity recognition task, and they can be extracted easily. The features are directly derived from the training corpus for training the machine learning model (or active learner in our case) and from the test corpus. The trained machine learning model can predict appropriate labels based on the extracted features. The features extracted include features of context words, word prefix, and suffix, beginning and ending word of the sentence, word length, POS tag (using nltk library), word case, word type, word pattern, character n-grams for word and POS tag and lower case of the word. We could have included more features to increase the classifier's performance, but our focus is on evaluating the performance of the active learning approach for this experiment. The active learning-based approach is an iterative approach that takes a considerable

Table 2 Details of NCBI Disease corpus used in experiment

Characteristics	Train set	Development set	Test set	Total classes (excluding Others tag)
No. of abstracts	593	100	100	One class (Disease)
No. of sentences	5424	923	940	
No. of tokens	135,701	23,969	24,497	

Table 3 Details of BC5CDR corpus used in experiment

Characteristics	Train set	Development set	Test set	Total classes (excluding Others tag)
No. of abstracts	500	500	500	Two classes (Disease and Chemical)
No. of sentences	4560	4581	4797	
No. of tokens	118,170	117,453	124,750	

Table 4 Details of Spanish CoNLL 2002 corpus used in experiment

Characteristics	Train set	Development set	Test set	Total classes (excluding others tag)
No. of sentences	8323	1915	1517	Eight classes (Organization, Location, Person and Miscellaneous)
No. of tokens	264,715	52,923	51,533	

Table 5 Features extracted for a sentence with five tokens from the train set of NCBI Disease corpus

Sentence	Sjogren	-	Larsson	syndrome	
POS	NNP	:	NNP	NN	
Label	B-Disease	I-Disease	I-Disease	I-Disease	O
Features	{'bias': 1.0, 'word.lower()': 'sjogren', 'word.len()': 7, 'word[-4:]': 'gren', 'word[-3:]': 'ren', 'word[-2:]': 'en', 'word[:2]': 'Sj', 'word[:3]': 'Sjo', 'word[:4]': 'Sjog', 'word.type': 'Alpha', 'word.case()': 'Title', 'word.pattern()': 'ULLLLLL', 'postag': 'NNP', 'postag[:2]': 'NN', 'BOS': True, '+1:word.lower()': '-', '+1:word.len()': 1, '+1:word.type': 'None', '+1:word.case()': 'None', '+1:word.pattern()': '-', '+1:postag': ':', '+1:postag[:2]': ':'}	{'bias': 1.0, 'word.lower()': '-', 'word.len()': 1, 'word[-4:]': '-', 'word[-3:]': '-', 'word[-2:]': '-', 'word[:2]': '-', 'word[:3]': '-', 'word[:4]': '-', 'word.type': 'None', 'word.case()': 'None', 'word.pattern()': '-', 'postag': ':', 'postag[:2]': ':', '-1:word.lower()': 'sjogren', '-1:word.len()': 7, '-1:word.type': 'Alpha', '-1:word.case()': 'Title', '-1:word.pattern()': 'ULLLLLL', '-1:postag': 'NNP', '-1:postag[:2]': 'NN', '+1:word.lower()': 'larsson', '+1:word.len()': 7, '+1:word.type': 'Alpha', '+1:word.case()': 'Title', '+1:word.pattern()': 'ULLLLLL', '+1:postag': 'NNP', '+1:postag[:2]': 'NN'}	{'bias': 1.0, 'word.lower()': 'larsson', 'word.len()': 7, 'word[-4:]': 'sson', 'word[-3:]': 'son', 'word[-2:]': 'on', 'word[:2]': 'La', 'word[:3]': 'Lar', 'word[:4]': 'Lars', 'word.type': 'Alpha', 'word.case()': 'Title', 'word.pattern()': 'ULLLLLL', 'postag': 'NNP', 'postag[:2]': 'NN', '-1:word.lower()': '-', '-1:word.len()': 1, '-1:word.type': 'None', '-1:word.case()': 'None', '-1:word.pattern()': '-', '-1:postag': ':', '-1:postag[:2]': ':', '+1:word.lower()': 'syndrome', '+1:word.len()': 8, '+1:word.type': 'Alpha', '+1:word.case()': 'Lower', '+1:word.pattern()': 'LLLLLLLL', '+1:postag': 'NN', '+1:postag[:2]': 'NN'}	{'bias': 1.0, 'word.lower()': 'syndrome', 'word.len()': 8, 'word[-4:]': 'rome', 'word[-3:]': 'ome', 'word[-2:]': 'me', 'word[:2]': 'sy', 'word[:3]': 'syn', 'word[:4]': 'synd', 'word.type': 'Alpha', 'word.case()': 'Lower', 'word.pattern()': 'LLLLLLLL', 'postag': 'NN', 'postag[:2]': 'NN', '-1:word.lower()': 'larsson', '-1:word.len()': 7, '-1:word.type': 'Alpha', '-1:word.case()': 'Title', '-1:word.pattern()': 'ULLLLLL', '-1:postag': 'NNP', '-1:postag[:2]': 'NN', '+1:word.lower()': '.', '+1:word.len()': 1, '+1:word.type': 'None', '+1:word.case()': 'None', '+1:word.pattern()': '.', '+1:postag': '.', '+1:postag[:2]': '.'}	{'bias': 1.0, 'word.lower()': '.', 'word.len()': 1, 'word[-4:]': '.', 'word[-3:]': '.', 'word[-2:]': '.', 'word[:2]': '.', 'word[:3]': '.', 'word[:4]': '.', 'word.type': 'None', 'word.case()': 'None', 'word.pattern()': '.', 'postag': '.', 'postag[:2]': '.', '-1:word.lower()': 'syndrome', '-1:word.len()': 8, '-1:word.type': 'Alpha', '-1:word.case()': 'Lower', '-1:word.pattern()': 'LLLLLLLL', '-1:postag': 'NN', '-1:postag[:2]': 'NN', 'EOS': True}

amount of time to finish the experiment, especially for a big corpus like JNLPBA. These common and simple features help in the execution of the experiment in an appropriate timeframe. The example of the features extracted for a short sentence from the NCBI disease corpus training set is shown in Table 5.

4.2 Baseline classifier

The conditional random field (CRF) classifier is considered and used as a baseline classifier for named entity recognition task. We have also employed a CRF classifier for the active learning-based approach used in this experiment. The python-based sklearn-crfsuite is a python-crfsuite wrapper used to execute the CRF classifier in the proposed algorithm. It also provides other utilities, such as metrics for evaluation. As a training algorithm, we have used L-BFGS algorithm with an elastic net, i.e., with L1 and L2 regularization. We have followed [35] for setting the other parameter values of the algorithm.

4.3 Evaluation metrics

Whether active learning or supervised learning, F-score is used for evaluation of the learner, which is the harmonic mean of the recall and precision. For multiclass classification, F-score is a usually considered preferable evaluation metric as it works well for imbalanced data, which is often the case in sequence labeling tasks. The recall and precision are calculated using the following performance measures: true positive, true negative, false positive, and false negative. A true positive outcome occurs when the classifier correctly predicts the positive label. Similarly, true negative outcome occurs when the classifier rightly predicts the negative label. A false positive outcome occurs when the classifier predicts the positive label wrongly. Similarly, false negative outcome occurs when the classifier predicts the negative label wrongly [36]. The evaluation metrics can be given as follows [33]:

$$\text{Precision} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Positive}}$$

$$\text{Recall} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Negative}}$$

$$F - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5 Active learning

5.1 Active learning approach

Active learning is one of the techniques that belong to machine learning. The active learner can sample the most appropriate data instances from the unlabeled data pool, which are further added to the labeled train set after being labeled by the oracle. The oracle is often a human expert in the respective domain. The active learning-based approach can achieve higher accuracy with considerably less labeled training instances than a supervised or passive learning approach [37]. In the active learning approach, whole train set is unlabeled and the train set is divided into two sets, i.e., the labeled set L and the unlabeled set U. Initially, the labeled set L is empty, and the whole train data are present in the unlabeled set U. The active learning approach begins by randomly selecting a few seed sentences to train the active learner. These randomly selected seed sentences are labeled by the oracle and placed in labeled train set L that was initially empty. The active learner is trained over the labeled train set L, and the trained active learner predicts the probability for each word of every sentence present in the unlabeled set U. An appropriate query sampling strategy is used by the active learner to select the most appropriate data instances, i.e., sentences in our case. These selected sentences are again labeled by the oracle and then added to the train set L, and the above process repeats until the stop condition is met. The labeled training set L gets bigger with each iteration, and only appropriate sentences are added to this set after labeling such that the active learner is able to learn quickly. The above-discussed active learning approach can be better understood with the algorithm presented in Algorithm A and from Fig. 1 that presents the active learning process.

Algorithm A: General steps in active learning algorithm

1 : Initially, select few seed data instances randomly from unlabeled train set U
2 : Oracle labels the selected data instances and place it in labeled set L
3 : Train the classifier C using labeled set L
4 : For each data instances in unlabeled train set U use trained classifier C to predict probability for each label
5 : Select most informative data instances from unlabeled train set U using query sampling strategy
6 : If stop criteria not reached then go to Step 2 else stop

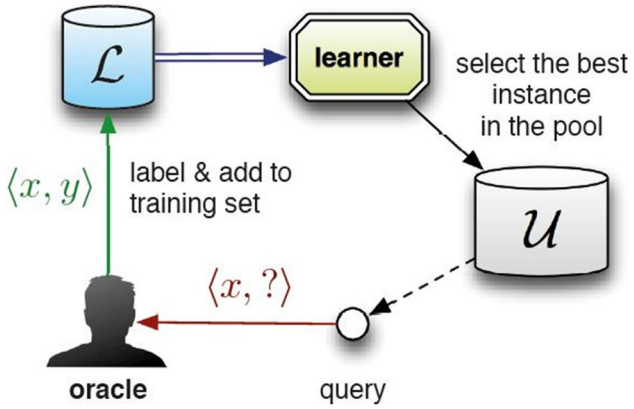


Fig. 1 Process in active learning approach (Source: [38])

5.2 Query sampling strategy: least confidence technique

The uncertainty-based sampling approach was initially used by [39] in which the learner selects the instances for which it is most uncertain. As the name suggests, the least confidence sampling approach selects the instance for which the active learner is least confident about its prediction for the most likely label. Earlier least confidence sampling approach is employed by [40] for sequence-based models where input sequence is denoted by x and label sequence is represented by y, and its query strategy formulation $\phi^{LC}(\cdot)$ can be written as follows [41]:

$$\phi^{LC}(x) = 1 - P(y^*|x;\theta) \quad (1)$$

where y^* is the most likely tag according to the learner. In this paper, we have proposed the least confidence query sampling strategy-based active learning approach for named entity recognition task. The proposed algorithm is presented in Algorithm 1 in Sect. 6.

5.3 Stop criteria

In the past, researchers have tried different stopping criteria to stop the active learning approach. The active learning approach can be stopped when there is no improvement in the performance with iteration [42] or after reaching user-set minimum absolute performance or maximum absolute performance [43], etc. According to [44, 45], in some realistic cases, when the labeled test set is present in advance, then the active learning approach can be stopped once it reaches a certain desired performance level. Therefore, in our experiment, we have first applied supervised learning over all the corpora using the same feature set and CRF classifier. The F-score is recorded for all the biomedical corpora and used as the desired performance level to stop the active learning approach. In short, it can be said that the active learning approach is stopped once it reaches the performance level attained by the supervised learning classifier for each of the corpora. The result obtained by the supervised CRF classifier using the above corpora and feature set is as follows in Table 6.

Table 6 Result obtained by supervised Conditional Random Field (CRF) classifier for different biomedical corpora

Corpus	Precision	Recall	F-score
JNLPBA dev set	80.3	76.3	78.1
JNLPBA test set	75.0	72.7	73.5
NCBI disease dev set	87.0	78.8	82.7
NCBI disease test set	85.7	78.2	81.8
BC5CDR dev set	85.0	73.9	79.1
BC5CDR test set	84.5	73.4	78.5
Spanish CoNLL 2002 dev set	77.9	75.4	76.3
Spanish CoNLL 2002 test set	80.8	79.2	79.8

5.4 Oracle

The oracle could be a human annotator or a labeling source [38] that provides the correct labels to the queried unlabeled data instances. The biomedical corpora considered for the experiment are already labeled by the experts, i.e., the words are labeled with correct labels for each of train set, development set, and test set. We have removed the labels of the train set so that it can be used as if it were a train set containing unlabeled text. Since we already had the correct labels for the train set, we assigned the correct labels to selected unlabeled sentences in place of the oracle.

6 Proposed active learning approach

The basic active learning algorithm is discussed in the previous section, which can be applied to different problems. In this paper, we have applied an active learning approach for named entity recognition problem using CRF classifier. As discussed above, we have used a python library, i.e., sklearn-crfsuite, to implement this task. The corpus is completely stored in the python list of list of tuples data structure such

that the outermost list contains inner lists which contain sentences. Each word of the sentences that is present in inner lists is stored in tuples along with labels. Before training the CRF model using the train set, the words inside the tuples are replaced with dictionary of features for the respective word. Similar replacement of words inside the tuples with a dictionary of features for each word takes place for the test set before prediction by the trained CRF classifier. So it is clear that the sentences are passed to train the model or active learner, not words. Keeping the above information in mind, the active learning-based algorithm is proposed for named entity recognition in which sentences are treated as specific instances. As discussed above, a query sampling strategy is used to select the most informative instances from the unlabeled train set U. In the proposed active learning-based approach for named entity recognition task, the least confident query sampling strategy is used, a widely used uncertainty query sampling strategy and discussed in the previous section. We have compared the proposed active learning approach that uses the least confident query sampling strategy with the active learning with other existing uncertainty-based query sampling strategies to evaluate their performance for named entity recognition task and compare the query sampling strategies with each other.

So in this paper, we have first implemented an active learning algorithm that adopts a random query sampling strategy by following general active learning algorithm steps discussed in Algorithm A for named entity recognition task. The algorithm is presented below in Algorithm B as follows.

Algorithm B: Active learning algorithm with random query sampling strategy for NER

1 : Select 1% of sentences randomly from train set to be used as seed data instances
2 : Oracle labels the selected sentences and place it in labeled set L
3 : Train the classifier C using sentences in labeled set L
4 : Select 10 sentences randomly from unlabeled train set U
5 : If stop criteria not reached then go to Step 2 else stop

Secondly, the existing active learning algorithm has been presented following a general active learning algorithm similar to the above algorithm that utilizes the least confidence query sampling strategy for named entity recognition and is presented as follows in Algorithm 1.

Algorithm 1: Existing active learning algorithm for NER using least confidence query sampling strategy

1 : Select 1% of sentences randomly from train set to be used as seed data instances
2 : Oracle labels the selected sentences and place it in labeled set L
3 : Train the classifier C using sentences in labeled set L
4 : For every sentences in unlabeled train set U:
  i. For each word of the sentence:
    a. Use trained classifier C to predict probability for each class label
    b. Compute least confidence score (lc_w) using probability value of most likely class label by applying equation (1)
  ii. Calculate the mean using all the above computed least confidence scores of words.
  iii. Store the calculated mean as the confidence score of the sentence.
5 : Sort the least confidence score of the sentences in increasing order and the sentence_id accordingly
6 : Select 10 most informative sentences from unlabeled train set U i.e., last 10 sentences having maximum least confidence scores from above sorted list
7 : If stop criteria not reached then go to Step 2 else stop

In the above existing algorithm, the mean of the least confidence score for all the words present in the sentences for the most likely class label is considered as the confidence score of the whole sentence. So it is clear that the final confidence score of the whole sentence depends on all the words present in the sentence. The main drawback of the existing algorithm is that all the words of the sentence are considered for the calculation of the confidence score of the sentence. The existing least confidence algorithm favors longer sentences due to which learner might not best possible instance from the unlabeled dataset [19]. For this research work, the idea is to consider the least confidence score of a few words (not all the words) to calculate the confidence score of the sentences.

In this paper, we have proposed a new algorithm that is similar to the first existing algorithm, but while computing the confidence score of the sentence, it can also consider the confidence score of one or more words present in the sentence. In this algorithm, we have calculated the mean of top N words having high least confidence scores within the sentence. Here, N can be equal to 1, 2, 3,…, and k depending on the algorithm variant and k is the length of sentence. For example, if the value of N is 3 in the algorithm, then the least confidence score for the sentence will be the mean of the top-3 words in the sentence for which the least confidence scores are maximum in comparison with the other words of the sentence. The proposed algorithm is presented in Algorithm 2 as follows.

The experiments have been conducted multiple times considering different values of N, i.e., 1, 2, 3, 4, 5, and k, and we have named the algorithm, respectively, as 'AL_LC_top-1′, 'AL_LC_top-2′, 'AL_LC_top-3′, 'AL_LC_top-4′, 'AL_LC_top-5,′ and 'AL_LC_top-k'. It is important to note that we have not calculated the confidence score for the sentences having a length less than N in the train set U where N is any one of 1, 2, 3, 4, 5 with the only exception of k and not considered them for selection by using the query strategy. Since the k is the length of the sentence in the top-k algorithm, it is different for sentences having different lengths and least confidence score is computed for each sentence in the train set U. Also, the algorithm presented in Algorithm 1 is equivalent to the algorithm proposed in Algorithm 2 with the value of N equal to k as both the algorithms consider the mean of least confidence score of all the words as the least confidence score of the sentence.

The proposed algorithm can be better understood from this sample calculation of the least confidence score for a sentence with the help of predicted probability values. A sample example for calculating the least confidence score for a sentence with hypothetical probability values that are assumed to be predicted by the active learner for three different labels is given for a better understanding of the proposed algorithm in Table 7.

Algorithm 2 : Proposed active learning algorithm for NER using modified least confidence query strategy

1 : Select 1% of sentences randomly from train set to be used as seed data instances

2 : Oracle labels the selected sentences and place it in labeled set L

3 : Train the classifier C using sentences in labeled set L

4 : For every sentences in unlabeled train set U:

i. For each word of the sentence
   a. Use trained classifier C to predict probability for each class label
   b. Compute least confidence score (lc_w) using probability value of most likely class label by applying equation (1)

ii. Sort the least confidence score of the words in increasing order and their index accordingly

iii. Use the last N least confidence scores for most likely class label of the words from above sorted list to calculate the least confidence score of the sentence. The least confidence score of sentence ($AL\_LC\_Top\_N$) is the mean of top N least confidence score within the sentence that can be calculated using the following formula:

$$AL\_LC\_Top\_N = \frac{\sum_{i=k-N+1}^{k} lc\_w_i}{N}$$

where N = any one of 1, 2, 3, ..., k according to algorithm and k is length of current sentence.

iv. Store above calculated mean as the least confidence score of the sentence

5 : Sort the least confidence score of the sentences in increasing order and the sentence_id accordingly

6 : Select 10 most informative sentences from unlabeled train set U i.e., last 10 sentences having maximum least confidence scores from above sorted list

7 : If stop criteria not reached then go to Step 2 else stop

Table 7 Sample example for calculation of least confidence score for a sentence having seven words

SENTENCE		w_1	w_2	w_3	w_4	w_5	w_6	w_7
LABEL A		0.9	0.95	0.4	0.7	0.08	0.24	0.01
LABEL B		0.01	0	0.5	0.1	0.2	0.6	0.02
LABEL C		0.09	0.05	0.1	0.2	0.72	0.16	0.97
LEAST CONFIDENCE SCORE OF WORDS (*lc_w*)		0.1	0.05	0.5	0.3	0.28	0.4	0.03
SORTED ORDER (*index i*)		③	②	⑦	⑤	④	⑥	①
LEAST CONFIDENCE SCORE OF A SENTENCE USING PROPOSED ALGORITHM ($AL\_LC\_Top\_N = \frac{\sum_{i=k-N+1}^{k} lc\_w_i}{N}$)	AL_LC_Top_1 (here N=1)	⑦ = 0.5						
	AL_LC_Top_2 (here N=2)	$\frac{⑦+⑥}{2} = 0.45$						
	AL_LC_Top_3 (here N=3)	$\frac{⑦+⑥+⑤}{3} = 0.4$						
	AL_LC_Top_4 (here N=4)	$\frac{⑦+⑥+⑤+④}{4} = 0.37$						
	AL_LC_Top_5 (here N=5)	$\frac{⑦+⑥+⑤+④+③}{5} = 0.316$						
	AL_LC_Top_k (here N=k = 7)	$\frac{⑦+⑥+⑤+④+③+②+①}{7} = 0.237$						

In the above example, it is assumed that there is a sentence having seven words. Each word can belong to any one class, i.e., can have anyone label, either label A or label B or label C. In place of step 4.i.a. of Algorithm 2, it is assumed that the hypothetical predicted probability values by the active learner are also given for each of the words of our example (in row 2–3 of Table 7). According to the active learner, the cell border of the predicted probability value of the most likely class label is highlighted by bolding the cells for each word of the sentence. The next step is to calculate the least confidence score for each of the words (i.e., *lc_w*) in the sentence using the maximum least confidence score for the most likely class label present in highlighted cells using Eq. (1) (step 4.i.b. in Algorithm 2 and row 4 in Table 7). Further, the least confidence scores are sorted out in increasing order, and the sorted order is represented by index *i* (in row 5 of Table 7). So in the above example, $i=7$ represents the index of maximum least confidence score of the words (*lc_w*) within the sentence, and $i=1$ represents the minimum least confidence score of the words (*lc_w*) within the sentence. The least confidence score of the sentence is calculated according to the value of the N (where N is equal to any one of 1, 2, 3, 4, 5, k, where k is sentence length, which is 7 for the example sentence discussed above). The mean of the N maximum least confidence scores of the words in the sentence takes part in the computation of the least confidence score of the sentence. It is clear from the above example that the value of the least confidence score for the sentences reduces with the increase in the value of N. The main reason is that more than one words take part in the calculation of the least confidence score of the sentence, not only the word having maximum least confidence score. It is important to note that the separate experiment is conducted for different values of N, and the value of N remains constant during a particular experiment. The idea of the above proposed algorithm is inspired by the normalized version of the least confidence query sampling strategies used by [19].

7 Results and discussion

In this paper, a new active learning approach is proposed for named entity recognition, which is a sequence labeling task. The proposed approach makes use of the least confidence query sampling strategy in which the different number of most uncertain words are considered for calculation of the final least confidence score of a sentence for comparison with the final least confidence score of the other sentences. The proposed approach and algorithm are discussed in detail in the previous section. For the proposed algorithm, experiment is conducted for different values of N, i.e., 1, 2, 3, 4, 5, and k, where k is the length of the sentence. The proposed approach is compared with the active learning approach with a random query sampling strategy. For fair judgment, the proposed approach is also compared with the active learning approach based on the marginal sampling strategy with the CRF model proposed by [18] and the entropy query sampling strategy [17, 20]. For comparison to be fair, the same features are extracted for all the above approaches. As discussed above, we have stopped all the active learning

Table 8 Total number of iterations and sentences required by various active learning-based approaches to reach the performance of supervised approach

S. no	Corpus	Approach	Number of seed sentences for initial training (1% of total training sentences)	F-score for dev set by supervised classifier	No. of iteration to reach to the performance of supervised classifier for dev set	Total no. of sentence used for training to reach to the performance of supervised classifier for dev set	F-score for test set by supervised classifier	No. of iterations to reach the performance of supervised classifier for test set	Total no. of sentences used for training to reach the performance of supervised classifier for test set
1	NCBI Disease Corpus	Random	54	82.69	380	3844	81.75	435	4394
		AL_LC_top-1			81	854		**85**	**894**
		AL_LC_top-2			**65**	**694**		92	964
		AL_LC_top-3			125	1294		114	1184
		AL_LC_top-4			116	1204		121	1254
		AL_LC_top-5			100	1044		104	1084
		AL_LC_top-k			117	1214		113	1174
		Marginal [18]			87	914		96	1004
		Entropy [20]			129	1334		100	1044
2	BC5CDR Corpus	Random	45	79.06	418	4215	78.49	397	4005
		AL_LC_top-1			311	3145		**172**	**1755**
		AL_LC_top-2			261	2645		215	2185
		AL_LC_top-3			271	2745		199	2025
		AL_LC_top-4			**233**	**2365**		212	2155
		AL_LC_top-5			268	2715		190	1935
		AL_LC_top-k			258	2615		236	2395
		Marginal [18]			267	2705		258	2615
		Entropy [20]			311	3145		180	1835
3	JNLPBA Corpus	Random	186	78.06	1037	10,546	73.46	1411	14,286
		AL_LC_top-1			562	5796		717	7346
		AL_LC_top-2			665	6826		657	6746
		AL_LC_top-3			722	7396		645	6626
		AL_LC_top-4			609	6266		652	6696
		AL_LC_top-5			**537**	**5546**		634	6516
		AL_LC_top-k			661	6786		**521**	**5386**
		Marginal [18]			642	5796		815	8326
		Entropy [20]			716	7336		734	7516

Table 8 (continued)

S. no	Corpus	Approach	Number of seed sentences for initial training (1% of total training sentences)	F-score for dev set by supervised classifier	No. of iteration to reach to the performance of supervised classifier for dev set	Total no. of sentence used for training to reach to the performance of supervised classifier for dev set	F-score for test set by supervised classifier	No. of iterations to reach the performance of supervised classifier for test set	Total no. of sentences used for training to reach the performance of supervised classifier for test set
4	Spanish CoNLL 2002 Corpus	Random	83	76.30	767	7743	79.82	631	6383
		AL_LC_top-1			194	2013		194	2013
		AL_LC_top-2			191	1983		**163**	**1703**
		AL_LC_top-3			220	2273		169	1763
		AL_LC_top-4			**169**	**1763**		284	2913
		AL_LC_top-5			175	1823		276	2833
		AL_LC_top-k			247	2543		318	3253
		Marginal [18]			252	2593		274	2813
		Entropy [20]			191	1983		313	3203

approaches when it reaches the performance of the supervised approach. For appropriate comparison among the active learning approaches, we have recorded the total number of iteration and the total number of sentences required to reach the performance level of the supervised approach for all the biomedical corpora considered in the experiment. The result of the above active learning-based approaches is presented in Table 8. The best results are highlighted in bold for the development set (dev) and the test set of each corpus, respectively.

For NCBI Disease corpus, the best result for the dev set is obtained by the proposed approach with $N=2$ (AL_LC_top-2), i.e., when the two maximum least confidence score of the words within the sentences are considered for calculation of the least confidence score of the sentence for comparison with least confidence score of the other sentences which is calculated in the same way. It took a total of 65 iterations and 694 sentences to reach the performance level of the supervised approach. The AL_LC_top-1 and marginal sampling follow the above algorithm, and they took 81 and 87 iterations, respectively, to reach the stop criteria. For the test set, the proposed approach with $N=1$ (AL_LC_top-1) performs best and is followed by AL_LC_top-2, the marginal sampling, and the entropy sampling strategy. So the proposed approach performs well for NCBI corpus with small values of N, i.e., 1 and 2. The active learning approach with random query sampling strategy takes the maximum iterations and number of sentences to reach the stop criteria. Also, the proposed approach with other value of N follows the above three algorithms to complete the task and reach to the stop criteria.

In case of the BC5CDR corpus, for dev set the proposed approach with $N=4$ (AL_LC_top-4) requires the minimum number of sentences to reach the stop criteria. The proposed approach with $N=$ k, 2, marginal sampling strategy, and other remaining approaches follows its performance to reach the stop criteria. Similarly, for the test set, the minimum number of sentences for training is needed by the proposed approach with $N=1$ (AL_LC_top-1), and it is followed by entropy sampling, the proposed approach with $N=5$, 3, 4, 2, and others. The active learning approach with random query strategy took the maximum number of sentences to reach the stop criteria.

For the JNLPBA dev set, the proposed active learning approach with $N=5$ performed best and reached the stop criteria in just 537 iterations. This approach is followed by the proposed algorithms with values of $N=1$, 4, and marginal query sampling strategy along with other approaches. For its test set, proposed approach with $N=$ k performed best, followed by $N=5$, 3, 4, and remaining approaches. Again the active learning approach with a random sampling strategy took the maximum number of iterations to reach the stop criteria.

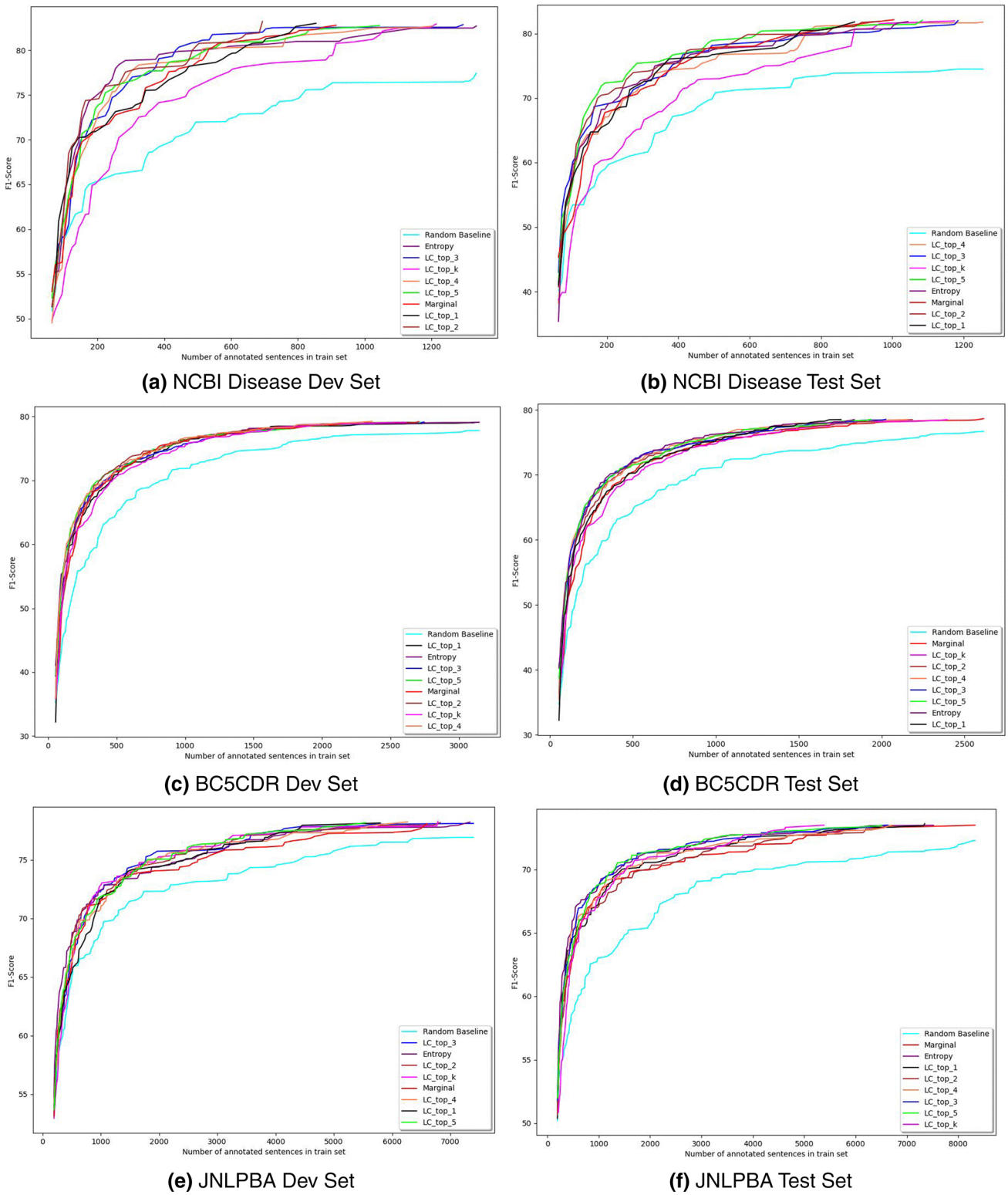
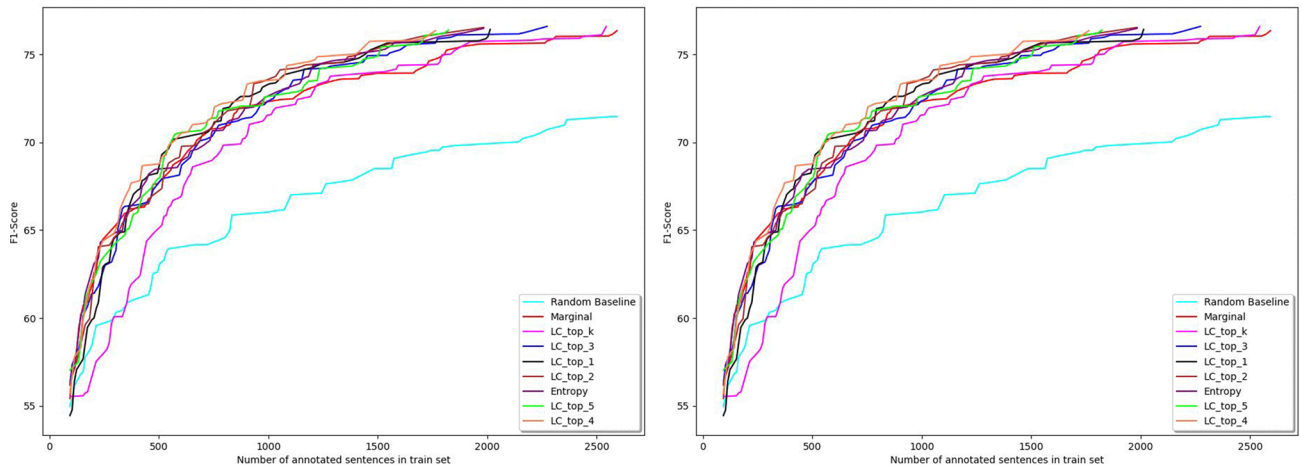


Fig. 2 Comparison of the performance of various active learning approaches over three biomedical corpora and a Spanish language corpus

For the Spanish CoNLL 2002 dev set, the proposed active learning approach with $N = 4$ and 5 performed the best. Their performance is followed by an entropy sampling strategy, the proposed approach with $N = 2$, 1, 3, k, and remaining others. For the Spanish CoNLL 2002 test set, the proposed approach with $N = 2$, 3, 1 performed best and reached the best result



(g) Spanish CoNLL 2002 Dev Set **(h)** Spanish CoNLL 2002 Test Set

Fig. 2 (continued)

in 163 to 194 iterations only. Their performance is followed by marginal sampling, the proposed approach with $N = 5$, 4, and entropy query sampling.

The above result is followed by the comparison plot in Fig. 2 of the performance of the active learning approaches. The x-axis represents the number of annotated sentences in the training set, and the y-axis represents the f-score accordingly. Graphs have been plotted for a total of eight active learning approaches, and it is found that they overlap over each other. So to compare the performance and to visualize the plots clearly, we have plotted the graph only with incremental F-score values in the y-axis with the number of annotated sentences in the training set on the x-axis.

Finally, the above active learning approaches have been compared with each other on the grounds of the percentage of labeled data (i.e., sentences) required to reach the stop criteria. The comparison is made between the active learning approaches, i.e., proposed approach (best result with suitable N), marginal sampling, entropy sampling, and random sampling approach. The comparison is presented in Fig. 3.

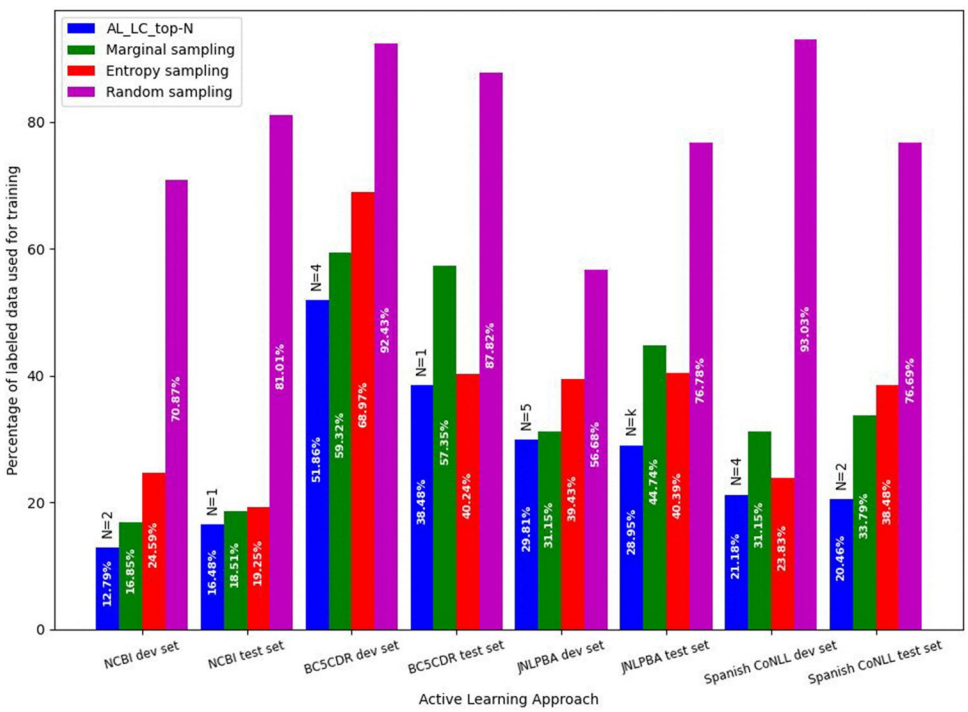


Fig. 3 Amount of data required for training by different query sampling strategies of the active learning approach over the dev set and test set of three different biomedical corpora and a Spanish language corpus

Observations made from the above result include the following:

- All the active learning approaches including the approach with a random sampling strategy reach the stop criteria, i.e., they always reach the performance level of the supervised classifier that is trained with all the labeled sentences present in the train set.
- Active learning approach with a random sampling approach performs poorly and requires about 55% to 93% of the training set to reach the stop criteria for the above corpora.
- It is clear from the result that the active learning approach with suitable query sampling strategy can enable the learner to reach appropriate performance level while requiring much less labeled data instances in the training set.
- For large corpus like JNLPBA, smaller values of *N* are preferable for the proposed approach. After considering the result, it is difficult to determine any one or two suitable values of N for the proposed approach for smaller corpus like NCBI Disease and BC5CDR corpus.
- In comparison with active learning approach with other sampling strategies, the proposed approach has the ability to perform better with suitable value of N, which has to be determined empirically.

8 Conclusions

In this paper, an efficient query sampling strategy was presented for the active learning algorithm of named entity recognition task. The proposed query sampling technique is based on modified least confidence sampling, which is an uncertainty query sampling strategy. It considers one or more most uncertain words (i.e., value of *N*) within the sentence for the calculation of the least confidence score of the sentence for comparison with the least confidence score of the other sentences. We have compared the proposed query strategy with the random sampling strategy and two other well-known existing uncertainty query strategy, to evaluate whether the proposed approach is effective. For the experiment, features extracted, CRF classifier is kept common, and python-based sklearn-crfsuite was used. The proposed approach is evaluated over the development set and the test set of the three biomedical corpora of different domains and a Spanish language corpus having different classes. The effectiveness of the query sampling approaches for active learning algorithm for named entity recognition problem is measured by the amount of labeled data required for training to reach the performance level set using the stop criteria. It is found that the proposed approach reaches the above performance level with minimal requirement of the labeled data. There is slightly more requirement of labeled data for training to reach stop criteria by existing query sampling approaches. The random sampling approach required the maximum amount of labeled data for reaching the above performance level. The proposed approach is evaluated with different values *N* in the experiment, and to use it in any other sequence labeling task, there is the requirement to determine the value of *N*, i.e., the number of words that takes part in the calculation of the least confidence score of the sentence which has to done empirically. In the future, the above approach will be evaluated with other uncertainty-based query strategies.

References

1. PMC Repository Information. https://www.ncbi.nlm.nih.gov/pmc/. Accessed 03 Aug 2019.
2. Benajiba, Y., Rosso, P., Lyhyaoui, A.: Implementation of the ArabiQA question answering system's components. In: Proceedings of the 2nd Information Communication Technologies International Symposium Workshop on Arabic Natural Language Processing, ICTIS-2007, pp. 3–5. Fez, Morroco (2007).
3. Abdi, A., Hasan, S., Arshi, M., Shamsuddin, S.M., Idris, N.: A question answering system in hadith using linguistic knowledge. Comput. Speech Lang. (2019). https://doi.org/10.1016/j.csl.2019.101023
4. Trisedya, B.D., Weikum, G., Qi, J., Zhang, R.: Neural relation extraction for knowledge base enrichment. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 229–240. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1023.
5. Khalifa, M., Shaalan, K.: Character convolutions for Arabic named entity recognition with long short-term memory networks. Comput. Speech Lang. **58**, 335–346 (2019). https://doi.org/10.1016/j.csl.2019.05.003
6. Aguilar, G., Maharjan, S., López-Monroy, A.P., Solorio, T.: A multi-task approach for named entity recognition in social media data. CoRR. abs/1906.0 (2019).
7. Yeniterzi, R., Tür, G., Oflazer, K.: Turkish named-entity recognition. In: Oflazer, K., Saraçlar, M. (eds.) Turkish Natural Language Processing, pp. 115–132. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90165-7_6.
8. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. CoRR. abs/1812.0 (2018).
9. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investig. **30**, 3–26 (2007). https://doi.org/10.1075/li.30.1.03nad
10. Krishnan, V., Manning, C.D.: An effective two-stage model for exploiting non-local dependencies in named entity recognition. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 1121–1128. Association for Computational Linguistics, Sydney, Australia (2006). https://doi.org/10.3115/1220175.1220316.
11. Sang, K.T.E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, pp. 142–147. Association for Computational Linguistics (2003).
12. Kazama, J., Makino, T., Ohta, Y., Tsujii, J.: Tuning support vector machines for biomedical named entity recognition. In:

Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, vol 3, pp. 1–8. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1118149.1118150.
13. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. CoRR. abs/1707.0 (2017).
14. Zhao, Z., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., Wang, J.: Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Med. Genomics. **10**, 73 (2017). https://doi.org/10.1186/s12920-017-0316-8
15. Campos, D., Matos, S., Oliveira, J.L.: Biomedical named entity recognition: a survey of machine-learning tools. In: Sakurai, S. (ed.) Theory and Applications for Advanced Text Mining. IntechOpen, Rijeka (2012). https://doi.org/10.5772/51066.
16. Chang, K.H.: Explaining active learning queries (2017).
17. Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C., Xu, H.: A study of active learning methods for named entity recognition in clinical text. J. Biomed. Inform. 58, 11–18 (2015). https://doi.org/10.1016/j.jbi.2015.09.010.
18. Ekbal, A., Saha, S., Sikdar, U.K.: On active annotation for named entity recognition. Int. J. Mach. Learn. Cybern. **7**, 623–640 (2016). https://doi.org/10.1007/s13042-014-0275-8
19. Liu, M., Tu, Z., Wang, Z., Xu, X.: LTP: a new active learning strategy for bert-crf based named entity recognition (2020).
20. Huang, H., Wang, H., Jin, D.: A low-cost named entity recognition research based on active learning. Sci. Program. **2018**, 10 (2018). https://doi.org/10.1155/2018/1890683
21. Skeppstedt, M., Paradis, C., Kerren, A.: PAL: a tool for pre-annotation and active learning. J. Lang. Technol. Comput. Linguist. **31**, 91–110 (2017)
22. Klie, J.-C.: INCEpTION: interactive machine-assisted annotation. In: Proceedings of the First Biennial Conference on Design of Experimental Search and Information Retrieval Systems. p. 105 (2018).
23. Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pp. 5–9. Association for Computational Linguistics (2018).
24. Kholghi, M., Sitbon, L., Zuccon, G., Nguyen, A.: Active learning reduces annotation time for clinical concept extraction. Int. J. Med. Inform. **106**, 25–31 (2017). https://doi.org/10.1016/j.ijmedinf.2017.08.001
25. Van Tran, C., Nguyen, T.T., Hoang, D.T., Hwang, D., Nguyen, N.T.: Active learning-based approach for named entity recognition on short text streams. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) Multimedia and Network Information Systems, pp. 321–330. Springer, Cham (2017)
26. Tran, V.C., Hoang, D.T., Nguyen, N.T., Hwang, D.: A hybrid method for named entity recognition on tweet streams. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) Intelligent Information and Database Systems, pp. 258–268. Springer, Cham (2017)
27. Project, G.: BioNLP/JNLPBA Shared Task 2004. http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004.
28. Collier, N., Kim, J.-D.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications ({NLPBA}/{B}io{NLP}), pp. 73–78. COLING, Geneva (2004).
29. Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics **19**, i180–i182 (2003)
30. Crichton, G., Pyysalo, S., Chiu, B., Korhonen, A.: A neural network multi-task learning approach to biomedical named entity recognition. BMC Bioinform. **18**, 368 (2017). https://doi.org/10.1186/s12859-017-1776-8
31. Doğan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform. **47**, 1–10 (2014). https://doi.org/10.1016/j.jbi.2013.12.006
32. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database. 2016, (2016). https://doi.org/10.1093/database/baw068.
33. Bhasuran, B., Murugesan, G., Abdulkadhar, S., Natarajan, J.: Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. J. Biomed. Inform. **64**, 1–9 (2016). https://doi.org/10.1016/j.jbi.2016.09.009
34. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of the 6th Conference on Natural Language Learning, vol 20, pp. 1–4. Association for Computational Linguistics, Stroudsburg (2002). https://doi.org/10.3115/1118853.1118877.
35. Korobov, M.: sklearn-crfsuite docs. https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html. Accessed 04 Nov 2019.
36. Classification: True vs. false and positive vs. negative. https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative.
37. Settles, B.: From theories to queries: active learning in practice. Active Learning and Experimental Design workshop In conjunction with AISTATS **2010**, 1–18 (2011)
38. Settles, B.: Active learning. Synth. Lect. Artif. Intell. Mach. Learn. **6**, 1–114 (2012). https://doi.org/10.2200/S00429ED1V01Y201207AIM018
39. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Cohen, W.W., Hirsh, H. (eds.) Machine Learning Proceedings 1994, pp. 148–156. Morgan Kaufmann, San Francisco (1994). https://doi.org/10.1016/B978-1-55860-335-6.50026-X.
40. Culotta, A., McCallum, A.: Reducing Labeling effort for structured prediction tasks. In: Proceedings of the 20th National Conference on Artificial Intelligence, vol 2, pp. 746–751. AAAI Press, Palo Alto (2005).
41. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1070–1079. Association for Computational Linguistics, Stroudsburg (2008).
42. Lin, Y., Sun, C., Xiaolong, W., Xuan, W.: Combining Self Learning and Active Learning for Chinese Named Entity Recognition. J. Softw. 5, (2010). https://doi.org/10.4304/jsw.5.5.530-537.
43. Laws, F., Schätze, H.: Stopping criteria for active learning of named entity recognition. In: Proceedings of the 22Nd International Conference on Computational Linguistics, vol 1, pp. 465–472. Association for Computational Linguistics, Stroudsburg (2008).
44. Vlachos, A.: A stopping criterion for active learning. Comput. Speech Lang. **22**, 295–312 (2008). https://doi.org/10.1016/j.csl.2007.12.001
45. Confidence-based active learning: Mingkun Li, Sethi, I.K. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1251–1261 (2006). https://doi.org/10.1109/TPAMI.2006.156