Clustering Based Active Learning for Biomedical Named Entity Recognition

Xu Han
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
Email: hanx0017@e.ntu.edu.sg

Chee Keong Kwoh
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
Email: asckkwoh@ntu.edu.sg

Jung-jae Kim
Institute for Infocomm Research
1 Fusionopolis Way
Singapore 138632
Email: jjkim@i2r.a-star.edu.sg

***Abstract*—The recognition and extraction of biomedical names is an essential task for the biomedical information extraction. However, the preparation of large annotated corpora hinders the training of the Named Entity Recognition (NER) systems. Active learning is reducing the needed manual annotation work in supervised learning task. In this work, we propose a novel clustering based active learning method for the biomedical NER task. We show that the underlying NER system using the proposed method outperforms those with other state of the art active learning methods, including density, Gibbs error and entropy based approaches, as well as the random selection. We compare variations of our proposed method and find the optimal design of the active learning method, which is to use the vector representation of named entities, and to select documents that are 'representative' and 'informative', as well as to use the Shared Nearest Neighbor (SNN) clustering approach. In particular, the optimal variant of the proposed method achieves a deficiency gain of 36.3% over the random selection.**

I. Introduction

Automatic recognition of biomedical names is an essential task in biomedical information extraction. While the training of many Named Entity Recognition (NER) systems follows the supervised learning framework, the preparation of a large training corpus is often a prerequisite for such systems. However, such training data are usually manually annotated, where the annotation process is time-consuming and expensive. In addition, in biomedical domain, the annotation process requires domain expertise, which makes the manual annotation more difficult. There is thus the need of reducing the amount of annotated data that are required for supervised learning systems.

Active learning is a special case of semi-supervised machine learning in which it chooses only 'informative' documents so that the needed manual annotation work is reduced to a minimum, while the performance of the supervised learning system is less or even not undermined with the reduced training data set [1]. Active learning works iteratively as follows: Among the unlabeled documents, the most 'informative' examples, whose annotations are most beneficial to the training of the underlying supervised learning system, are selected out for annotation. Once the selected documents are manually annotated by human annotator, the annotated documents are added into the training data set. And the learning system is updated accordingly. Such a cycle of 'selection, annotation, updating' iterates until termination conditions (e.g. no more unlabeled data, no/little change of system performance) are met.

The main issues in active learning is the selection of documents for annotation. Regarding to this issue, the commonly used active learning framework can be roughly classified into two approaches: committee-based approach [2] and uncertainty-based approach [3]. The committee-based approach, based on a committee of classifiers, selects the documents whose classifications have the greatest disagreements among the classifiers and passes them to human experts for annotation. While the uncertainty-based approach is to label the most uncertain samples by using an uncertainty scheme, such as entropy [4], [5], Gibbs error criterion [6]. However, they do not consider if the selected samples are 'representative' or not, where non-representative samples such as outliers may cause overfitting.

To select 'representative' examples and avoid outliers, there are other approaches. One approach is to model the distribution of the data set. For instance, Settles and Craven [7] proposed a new method for active learning, called information density, in which the informativeness of a sample is weighted by its average similarity to all other samples. This method was designed to avoid spending too much time on annotating outliers, which have low density. Another approach is to utilize unsupervised clustering analysis. For instance, in [8], [9], [10], the authors pre-clustered all documents using the bag-of-words model for data representation and selected the cluster centroids for initial annotation. Note that in the sub-sequent document selection, they used information density measurement, rather than unsupervised clustering, to determine the selection of examples for annotation.

We propose a novel clustering based active learning method for the task of named entity recognition (NER), where the main three questions are considered: 1) what data representation model to use for the documents, 2) which clustering method to use and 3) how to select documents from clustering results.

In particular, our method clusters documents not by using the bag-of-words model, but by using candidate named entities found in the documents by an underlying classifier, thus reflecting the distribution of named entities. It groups doc-

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

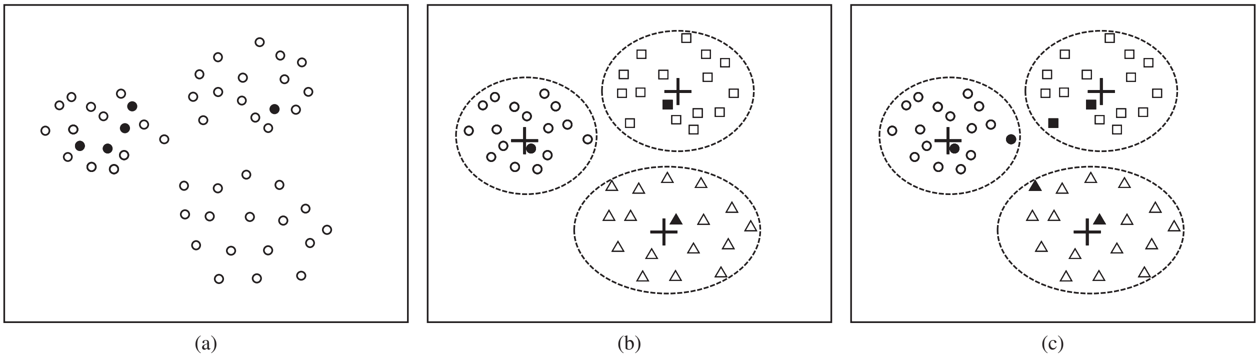

Fig. 1: Toy example demonstrating the selection strategies of density based and clustering based active learning method. Circles, triangles and squares denote the unlabeled samples, classified into different clusters. The crosses denote the supposed cluster centroid. Solid circles, triangles and squares denotes the selected samples for annotation. (a) Information Density approach selects documents that are within the dense areas. (b) Previous clustering method selects documents close to the cluster centroid. (c) Proposed method not only selects documents close to centroid, but also selects documents close to cluster boundary.

uments with *similar* named entities and selects for annotation those from different groups, on the assumption that similar named entities may share common features of the underlying classifier, thus redundant for training the classifier. We use a matrix factorization method [11] for the vector representation of named entities in order to measure the similarity of named entities. On the other hand, rather than merely selecting the cluster centroids for annotation, we select examples that are either cluster centroid or on cluster boundary, so that the selected examples are a mixture of 'representative' and 'uncertain' examples, which is expressed in Figure 1. Furthermore, we compare the variants of proposed method using different clustering methods.

This paper is organized as follows: Section 2 describes the related work. In Section 3 we discuss the method, followed by the experiment results and discussions in Section 4. Finally, Section 5 concludes this paper.

II. Related Work

A. Active learning

Active learning is a research topic in machine learning, and its main purpose is to reduce the amount of required training data in the supervised learning framework while also maintaining the performance of the supervised system. In active learning, the manual annotation work is kept to a minimum by choosing 'informative' documents for manual annotation such that the would-be annotations on the documents may promote the training of supervised learning systems more effectively than the other documents [1]. It has been studied for many natural language processing applications, such as word sense disambiguation [12], named entity recognition [13], [14], [15], speech summarization [16] and sentiment classification [17], [18].

Its existing works can be roughly classified into two approaches of uncertainty-based approach [3] and committee-based approach [2], as aforementioned. Entropy is a popular method for evaluating the uncertainty of examples [4], [5]. For the selection of examples, the larger the entropy of an example is, the more information the example is supposed to contain. An alternative measure is the Gibbs error criterion [6], and [6] shows that the Gibbs error criterion is superior to the entropy.

Recently, there are also research work that incorporate the support vector machine (SVM) into active learning, such as [19], [20], [21]. In these pieces of work, the samples that are close to the hyperplane (or decision boundary), called support vectors, are often considered as more important for classification than those that are far from the hyperplane. Therefore, in the SVM based active learning, those document that are close to the decision boundary are often selected as 'informative'.

It has been shown, however, that the uncertainty-based approach may have worse performance than random selection [22], [23], [24]. One limitation in the uncertainty-based approach is that they do not consider if the selected samples are 'representative' or not, where non-representative samples such as outliers may undermine the training performance of the supervised learning framework.

To tackle this issue, there are proposed active learning approaches that select 'representative' examples and avoid outliers. For instance, there is work of using the information density to guide the selection of documents [7]. This approach is to explicitly model the distribution of the data set with density weights, and is shown to outperform the uncertainty-based approaches On the other hand, there are also active learning methods that integrate the unsupervised clustering techniques [8], [9], [10]. For instance, in [8] they first cluster all samples before active learning and select the documents close to the cluster centroids as the initial set of annotation, instead of beginning with a random subset of samples. In [9] they further develop this idea by incorporating the information density to guide the document selection, through maintaining a k-nearest-neighbor-based density weight, in order to circumvent the risk of selecting outliers.

In this work, we integrate the two approaches by selecting a mixture of both 'representative' and 'informative' documents from clustering results. Different from [9], we consider the distance difference with respect to the centroids, rather than using the density weight.

B. Document representation model

Previous work of text clustering often adopt bag-of-word model, in which a document is represented as the bag of its words, discarding other properties of text, such as word order and grammatical structure. Term frequency-inverse document frequency (TF-IDF) is then used to identify word weights (or features) in each document and represent a document as a vector based on the word features [25]. In this work of active learning for NER, we group documents that contain similar named entities, on the assumption that the documents that share similar named entities may have common features of the underlying classifier, thus redundant for the manual annotation work.

C. Cluster analysis

The cluster analysis is an unsupervised task which is to group a set of data in such a way that the individual data in the same cluster are more similar to each other than data that belong to different clusters. The clustering has a long history and has been developed in domains such as statistics, pattern recognition, data mining, and other fields [26]. A large number of clustering techniques have been proposed, such as hierarchical clustering, centroid-based clustering, as well as density-based clustering [27], [28].

The hierarchical clustering seeks to build a hierarchy of clusters, and the strategies may be divided into agglomerative methods, which are a series of successive fusions of individual objects into clusters, and divisive methods, which partition the set of objects successively into finer groups. We use hierarchical agglomerative clustering (HAC) in our analysis, as implemented in ELKI toolkit [29]. In the K-means clustering of centroid-based clustering, the general idea is to find the k cluster centers and assign the objects to their nearest cluster center, such that the distances between the objects and cluster centers are minimized. While the K-means clustering is an NP-hard problem, there is work to optimize the speed of execution [30]. While in density-based clustering, the clusters are defined as areas of higher density than the remainder of the data set [31]. The popular density-based clustering methods are DBSCAN [32] and Shared Nearest Neighbor (SNN) clustering [33], as well as Greedy Variance Minimization (GVM) [1].

III. Method

Our method measures the informativity of unlabeled documents using a state-of-the-art NER system for the dataset. It works iteratively as follows: 1) We train the NER system based on initial training data. 2) We apply the trained NER system to all the unlabeled documents to identify candidate named entities (NEs). 3) We obtain the vector representation of each candidate NE by adopting a matrix factorization method [11] and 4) cluster the unlabeled documents with the vector representation of candidate NEs. 5) We select the mixture of most 'representative' and 'uncertain' documents in each cluster for the manual annotation and 6) update the NER system with the annotated documents.

Figure 2 illustrates the proposed method in pseudo codes. In the following subsections, we discuss about three questions, which are: 1) how to represent documents, 2) which documents to select for manual annotation among the results of document clustering, and 3) which clustering method to use. Note that the three questions are related to the steps (4) and (5) in the previous paragraph and that the implementation details of the other steps can be found in Section IV-A.

A. Vector representation model for documents

We group documents that contain similar named entities into the same class to reduce the manual annotation work. We represent the documents in the perspective of its contained named entities, in the form of vectors. For this purpose, we utilize GloVe [11], which analyzes latent semantics of words by using global matrix factorization and local context window methods. We train a word vector model with the entire MEDLINE corpus for the BioCreative task [34].

We estimate the vector of a document as the average of vectors of named entities the document contains. Equation (1) shows how to represent a document (D) as a vector with the named entities (n_i) found in the document, where $|NE(D)|$ indicates the number of named entities found in document D. Note that $\vec{D}$ is the vector representation of document D.

$$\vec{D} = \frac{1}{|NE(D)|} \sum_{\forall n_i \in D} \vec{n_i} \quad , if|NE(D)| > 0 \tag{1}$$

For multi-word terms as named entities, we simplify the estimation of their vectors as follows, where len(n_i) indicates the number of words in the named entity n_i, and w_k indicates the kth word in n_i, and $\vec{n_i}$ is the vector representation of named entity n_i.

$$\vec{n_i} = \frac{1}{len(n_i)} \sum_{\forall w_k \in n_i} \vec{w_k} \tag{2}$$

For documents that no NE is identified by the NER system (i.e., $|NE(D)|$=0), we use the average of vector of word contained in the document, as shown in Equation (3). The w_j indicates the jth word in document, and $w(D)$ indicates the number of words found in document D, the $\vec{w_j}$ is the vector representation of word w_j.

$$\vec{D} = \frac{1}{|w(D)|} \sum_{\forall w_j \in D} \vec{w_j} \quad , if|NE(D)| = 0 \tag{3}$$

B. Document selection strategy from clustering result

The main issue in active learning is the selection of documents for manual annotation. In clustering based active learning, there is a question of how to select the documents

[1] http://www.tomgibara.com/clustering/fast-spatial/

```
Input: labeled document pool $L$, unlabeled document pool $U$, batch size $b$
// Initialization
$NE_0$ = the set of named entities annotated on $L$
Learn an NER model $M_0$ from $NE_0$
$i$ = 0 // the index of the current round
// Active Learning Loop
while $U$ is not empty:
  $i$ += 1
  // 1. Data Representation
    for each document $D_{ij}$ in $U$:
      Apply $M_{i-1}$ to $D_{ij}$ and collect the resultant named entities set $NE_{D_{ij}}$
      Construct the vector representation $\vec{D_{ij}}$ using $NE_{D_{ij}}$, (Equations (1,3))
  // 2. Clustering
    Cluster all $D_{ij}$ of $U$ into $\frac{b}{2}$ clusters.
  // 3. Document Selection
    Based on clustering results, select the $b$ documents, (Equations (4,5)), designated as $B$, for annotation by annotator.
    Remove $B$ from $U$, add $B$ to $L$, and add the annotations on $B$ to $NE_{i-1}$, designated as $NE_i$
    Learn a new model $M_i$ from $NE_i$
```

Fig. 2: Proposed clustering based active learning method

based on document clustering result. In this work, we select both ‘representative’ documents and ‘informative’ documents as aforementioned.

By ‘representative’, we follow the previous work [9], which is to select documents that are close to the cluster centroid, that is, the document’s distance from the cluster centroid is the minimum among those distances of the other documents that are within the same cluster, as expressed in Equation (4). The d indicates the document that belongs to the cluster C, and x_C is the centroid for cluster C. $dist$ indicates the distance between the document and the centroid. The $\tilde{d}_{representative}$ stands for the resultant ‘representative’document that is selected.

$$\tilde{d}_{representative} = \underset{d}{argmin}\ dist(d, x_C) \quad \forall d \in C \tag{4}$$

By ‘informative’, we mean the documents that are close to the boundary of the cluster, as its distance from its cluster centroid is maximized, comparing to other documents from the same cluster. Note that such documents could also be outliers. To circumvent of selecting outliers, we further require that such documents are close to the cluster’s boundary with other clusters, not to the boundary with no neighbor cluster. In this way, we choose the document whose distance to the cluster centroid is maximized, yet the difference between its distances to the cluster centroid and the neighbor cluster centroid is minimized. We consider that the annotations of such documents are ‘informative’ for the clustering, and also the training of the underlying system. These selection criteria are expressed in Equation (5). $\tilde{C'}$ denotes a neighboring cluster for C, and $x_{\tilde{C'}}$ is the centroid for cluster $\tilde{C'}$. The $\tilde{d}_{informative}$ stands for the chosen document that meets both of the requirements. In cases when no document is found to meets the conditions, the selection of ‘informative’ document is skipped.

$$\tilde{d}_{informative} = \begin{cases} \underset{d}{argmax}\ dist(d, x_C) \quad \forall d \in C \\ \underset{d}{argmin}\ (dist(d, x_{\tilde{C'}}) - dist(d, x_C)) \\ \qquad\qquad \forall d \in C \end{cases} \tag{5}$$

To find the neighboring cluster $\tilde{C'}$, we calculate the distances between the cluster centroids, which is expressed in Equation (6).

$$\tilde{C'} = \underset{C'}{argmin}\ dist(x_{C'}, x_C) \tag{6}$$

We select a mixture of documents from each cluster that meet the requirements of Equations (4-5), one using Equation (4) and the other using Equation (5).

C. Clustering method for active learning

In this work, we propose the clustering based active learning method. Then there is the question of whether the different clustering results from the clustering methods would affect the performance of active learning methods, as well as to determine which clustering approaches is appropriate for the proposed method. We adopted different clustering methods, including the HAC [26], DBSCAN [32], SNN [33] and GVM for the analysis.

IV. Experiment Results

A. Datasets and NER systems

In this work, we use the dataset from the gene mention recognition task of BioCreative II [34], and we utilize the state-of-the-art NER system, Gimli [35], for training and evaluating our method.

Methods for comparison: Each experiment starts with a held-out labeled development dataset for initialization and a pool of unlabeled training dataset for selection, and continues with ten rounds. In each round, 10% of the documents in the training dataset are selected by different sample selection strategies. For evaluation, we report the performance of the Gimli trained with the selected training document in each round, against the held-out test dataset in official evaluation procedure.

The sample selection strategies are as follows:

- Random selection: We randomly split the training dataset into 10 bins in advance, and one bin is randomly chosen in each round. Following 10-fold cross validation, we report the averaged performance. (hereafter referred to as Random Selection)
- Entropy-based active learning: We follow the uncertainty-based active learning method, and use the entropy as the selection criteria. We calculate the entropy of the documents, and select the documents that have top entropy values. The calculation of entropy follows the sequence entropy, proposed in [7], and is designated as AL(Entropy).
- Maximum Gibbs Error based active learning: Similar to the entropy-based method, we select documents that have top values in the Gibbs error, as introduced in [6], designated as AL(Gibbs).
- Cluster and Density based active learning: As proposed in [9], we use clustering to select the initial documents and then select documents based on document density for sub-sequential rounds. This is designated as AL(Cluster+Density).
- Density based active learning: We simulate the selection process that is purely based on document density, what is, without first selecting the initial document set using clustering approach. This is a variant of the work in [9], and is designated as AL(Density).
- Proposed clustering based active learning method: We use the proposed clustering based active learning method, introduced in the section III. Note that this includes the experiment design for the three subsection questions, which is to use the vector representation model for document, and use the proposed document selection strategy from clustering results, as well as the SNN clustering method. This is designated as AL(proposed).

B. Evaluation metrics

To compare the performance of the different strategies of sample selection, we plot their performance after each of the ten rounds of iteration. Since the difference between some plots is not obvious, however, we also use the evaluation metric of *deficiency* [9], [36], defined as follows:

$$Def_n(AL, REF) = \frac{\sum_{t=1}^{n}(acc_n(REF) - acc_t(AL))}{\sum_{t=1}^{n}(acc_n(REF) - acc_t(REF))} \qquad (7)$$

, where acc_t is the performance of the underlying classifier at t^{th} round of learning iteration. AL is an active learning method and REF is the baseline method (i.e. RS_Average). n refers to the total number of rounds (i.e. 10). Deficiency value smaller than 1.0 means that the active learning method is superior to the baseline method, and vice versa: In general, a smaller value indicates a better method.

C. Results

We applied these methods to the BioCreative dataset, plotted the learning curve of Gimli in Figure 3, and summarized their deficiencies in Table I. As shown in Figure 3, our clustering method AL(proposed) outperforms the previous clustering method AL(Cluster+Density) of [9], always showing steady improvement. In contrast, AL(Cluster+Density) falters when 20% and 40% of samples are selected, which may mean that its density-based method requires significant amount of samples for further improvement of the underlying classifier. This is also shown in the AL(Density) when 10% and 40% of the documents are selected, where its learning curve is increasing slower than the rest of methods, which may be due to the lack of training data for this method. The curves of the generic uncertainty based methods (e.g. entropy, Gibbs error) do not show significant improvement over the random selection.

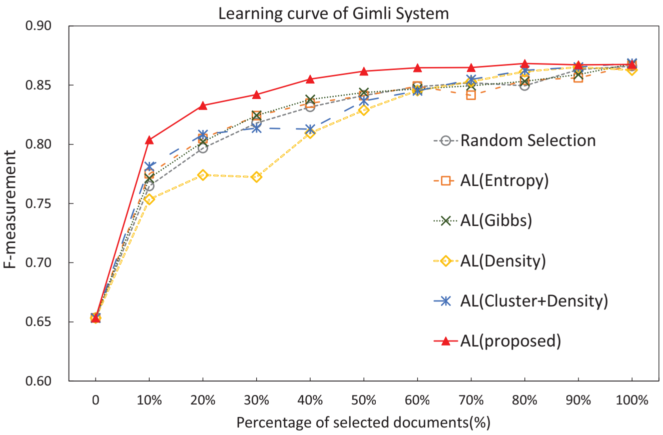

Fig. 3: Comparison of active learning methods and random selection with BioCreative dataset

In terms of deficiency, the AL(proposed) method gains a 36.3% improvement compared to the baseline, as shown in Table I. Note that the previous method AL(Density) shows higher deficiency than the baseline.

TABLE I: Comparison of deficiencies of acitve learning methods and random selection

Method	Deficiency
Random Selection	1
AL(Entropy)	0.972
AL(Gibbs)	0.954
AL(Density)	1.194
AL_Cluster+Density	0.968
AL(proposed)	**0.637**

1) Vector representation model for documents: In this part, we compare the effect of the vector representation in the proposed clustering based active learning method. Specifically, we compare the learning curves of the active learning method using different data representation models.

The compared document representation models are as follows:

– Vector representation of candidate NE: As proposed in Equations (1), (2) and (3), we use the vector representation of the recognized NEs in the document, designated as AL(VectorData_NE).
– Vector representation without prediction of candidate NE: Not considering the NEs recognized in the document, we only use the vector representation of all the words in the document (i.e. only using Equation (3)). This is designated as AL(VectorData_w/o NE).
– Bag-of-word model for documents: We follow the traditional bag-of-word model, which is to first create a feature set using the TF-IDF approach, and represent the documents using the feature set as vectors, which is designated as AL(Bag-of-Word).
– Random selection: This is the same as in the previous experiment, and is designated as Random Selection.

Considering the time efficiency, we carried out this experiment using the GVM clustering approach, and we always selected the first document in the clustering result, rather than selecting the documents close to the centroid or to the boundaries of clusters. The learning curves are plotted in Figure 4, and the deficiencies are summarized in Table II.

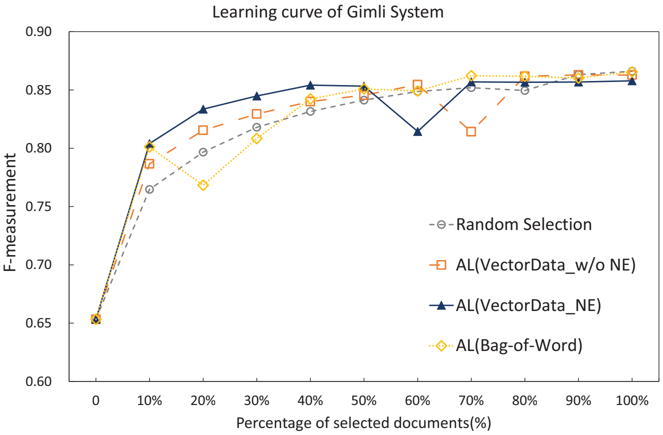

Fig. 4: Comparison of the performance of active learning methods with different representation models for documents, as well as the baseline of random selection. The vector representation (VectorData) is used either with or without prediction of candidate NE (NE). The performance of active learning using classicial bag-of-word model is also presented.

As shown in Figure 4, the learning curves using vector representation model are showing better performance than that of Bag-of-Word model. In particular, by using only the predicted NE rather than averaging over the whole sentence, the performance of the learning curve is further improved, which suggests that the process of prediction of NE helps to filter out redundant information in the sentence. Considering that in this experiment only the first document in the clustering result is selected, the learning curve is not showing steady improvement and affects from sudden drops.

TABLE II: Comparison of deficiencies of active learning methods with different data representation models for documents

Method	Deficiency
Random Selection	1
AL(Bag-of-Word)	0.929
AL(VectorData_w/o NE)	0.923
AL(VectorData_NE)	**0.814**

In terms of deficiency, the AL(VectorData_NE) method gains a 18.6% improvement compared to the baseline, as shown in Table II, while all the other methods show better deficiencies than the baseline.

2) Document selection strategy from clustering result: In this part, we compare the different document selection strategies from clustering result of active learning method. Concretely, we compare the following document selections strategies:

– First document in clustering result: We consider the speed efficiency and select the first document in the clustering result, skipping the calculation load for finding centroid. This is designated as AL(First in Cluster).
– Document closest to centroid: We follow the traditional approach that select the document closest to the cluster centroid, i.e., the documents that satisfy the Equation (4), which is designated as AL(ClusterCentroid).
– Proposed document selection strategy: We select the documents that meet the requirement in Equation (4) or Equation (5), and is designated as AL(ClusterCentroid+Edge).
– Document Scoring strategy: We select the document that has the top score in a classifier-independent scoring measurement, which is introduced in [37].
– Random selection: This is the same as in the previous experiment, and is designated as Random Selection.

We carried out this experiment using the GVM clustering approach, and we used the vector representation with the predicted NE, as introduced in Equation (1), Equation (2) and Equation (3). The learning curves are plotted in Figure 5, and the deficiencies are summarized in Table III.

As shown in Figure 5, the learning curves of the methods that select document closest to cluster centroid are better than that of random selection, while the proposed method performs even better. This becomes evident when about 30% to 70% of the documents are selected for manual annotation, where the learning curve of selecting only the documents close to centroid does not improve significantly. The strategy of selecting the first document in clustering results becomes less stable and drops when 60% of the documents are selected. In contrast, the selection strategy based on a scoring system

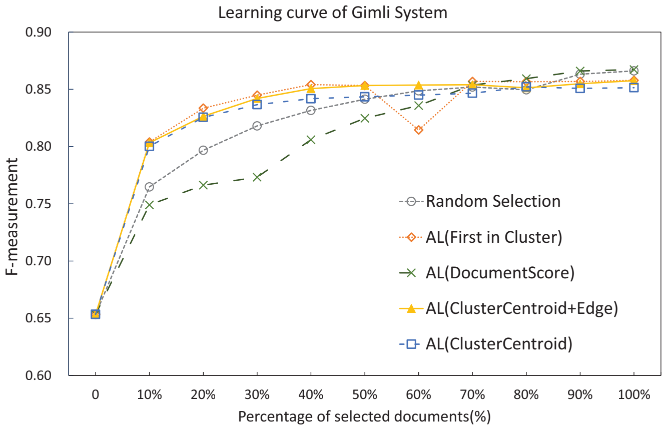

Fig. 5: Comparison of the performance of active learning methods with different document selection strategies, as well as the baseline of random selection.

performs worse than the random selection, which suggests that the scoring system may not be compatible with the clustering approach.

TABLE III: Comparison of deficiencies of active learning methods with different document selection strategies

Method	Deficiency
Random Selection	1
AL(First in Cluster)	0.814
AL(DocumentScore)	1.241
AL(ClusterCentroid)	0.886
AL(ClusterCentroid+Edge)	**0.787**

In terms of deficiency, the AL(ClusterCentroid+Edge) method gains a 21.3% improvement compared to the baseline, as shown in Table III, while the rest selection methods achieve better deficiencies than the baseline. However, the only exception is the scoring method, which shows worse deficiency than the random selection.

3) Clustering method for active learning: In this part, we compare how the different clustering methods affect the learning curves of active learning method. Particularly, we compared the clustering method of HAC, DBSCAN, SNN and GVM.

In this experiment, we use the vector representation of NE, introduced in Equation (1), Equation (2) and Equation (3), and we select documents using the strategy introduced in Equation (4) and Equation (5). We plot the learning curves in Figure 6 and compare the deficiencies in Table IV.

In Figure 6, the learning curve of HAC method is performing worse than that of random selection, while the rest method shows better performance. Particularly, the SNN approach shows the best performance for this experiment. One possible explanation for the robust performance of SNN is that such a method may better handle data of heterogeneous densities [33].

As shown in Table IV, the SNN method shows a 36.3% improvement compared to the baseline, while only the HAC method shows worse deficiency than that of random selection.

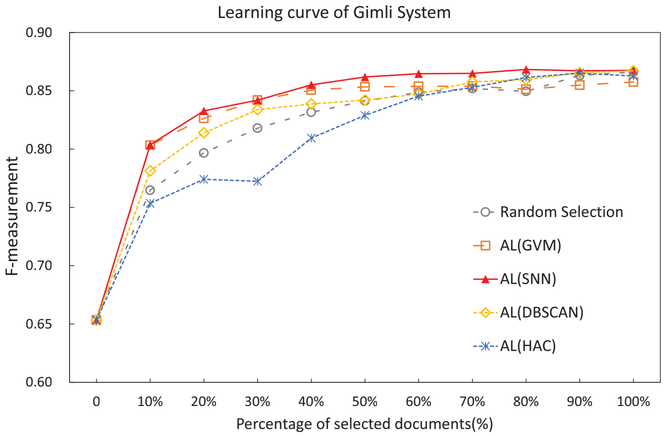

Fig. 6: Comparison of the performance of active learning methods with different clustering methods, as well as the baseline of random selection. For the clustering methods, the HAC, GVM, DVSCAN and SNN are considered.

TABLE IV: Comparison of deficiencies of active learning methods with different clustering methods

Method	Deficiency
Random Selection	1
AL(HAC)	1.194
AL(GVM)	0.787
AL(DBSCAN)	0.861
AL(SNN)	**0.637**

V. Conclusion

In this study, we propose a novel clustering based active learning method to select the documents that are 'informative' and 'representative'. Comparing to the previous clustering based active learning method, we show that the clustering of named entities surpass the benchmark of entropy, Gibbs error, and density based active learning methods, as well as the random selection. In addition, in terms of the deficiency gain, the proposed method can achieve a deficiency gain of 36.3% over the random selection. We also show that the performance of proposed active learning method is improved by using the vector representation of named entities in the documents, as well as adopting the SNN clustering approach.

Acknowledgment

This research was partially supported by Ministry of Education, Singapore, grant (MOE2014-T2-2-023).

References

[1] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, Jun. 2012.

[2] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.

[3] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the Eleventh International Conference on Machine Learning*, 1994, pp. 148–156.

[4] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "Entropy-based active learning with support vector machines for content-based image retrieval," in *Proceedings of IEEE International Conference on Multimedia and Exposition (ICME)*, 2004, pp. 85–88.

[5] A. Holub, P. Perona, and M. Burl, "Entropy-based active learning for object recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2008, pp. 1–8.

[6] N. V. Cuong, W. S. Lee, N. Ye, K. M. A. Chai, and H. L. Chieu, "Active learning for probabilistic hypotheses using the maximum gibbs error criterion," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 1457–1465.

[7] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 1070–1079.

[8] J. Kang, K. Ryu, and H.-C. Kwon, "Using cluster-based sampling to select initial training set for active learning in text classification," in *Advances in Knowledge Discovery and Data Mining*, 2004, pp. 384–388.

[9] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "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)*, August 2008, pp. 1137–1144.

[10] L. Qian and G. Zhou, "Clustering-based stratified seed sampling for semi-supervised relation classification," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, October 2010, pp. 346–355.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2014, pp. 1532–1543.

[12] Y. Chen, H. Cao, Q. Mei, K. Zheng, and H. Xu, "Applying active learning to supervised word sense disambiguation in medline," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 1001–1006, 2013.

[13] K. Tomanek and U. Hahn, "Reducing Class Imbalance During Active Learning for Named Entity Annotation," in *Proceedings of the Fifth International Conference on Knowledge Capture*, 2009, pp. 105–112.

[14] ——, "Semi-supervised active learning for sequence labeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, August 2009, pp. 1039–1047.

[15] ——, "A Comparison of Models for Cost-Sensitive Active Learning," in *Coling 2010: Posters*, Aug. 2010, pp. 1247–1255.

[16] J. Zhang and H. Yuan, "A Certainty-Based Active Learning Framework of Meeting Speech Summarization," in *Computer Engineering and Networking SE - 28*, 2014, vol. 277, pp. 235–242.

[17] S. Li, S. Ju, G. Zhou, and X. Li, "Active learning for imbalanced sentiment classification," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 139–148.

[18] S. Ju and S. Li, "Active learning on sentiment classification by selecting both words and documents," in *Chinese Lexical Semantics*, 2013, vol. 7717, pp. 49–57.

[19] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[20] X. Li and Y. Guo, "Active learning with multi-label svm classification." in *IJCAI*. Citeseer, 2013.

[21] J. Kremer, K. Steenstrup Pedersen, and C. Igel, "Active learning with support vector machines," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 4, pp. 313–326, 2014.

[22] H. Schütze, E. Velipasaoglu, and J. O. Pedersen, "Performance Thresholding in Practical Text Classification," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 2006, pp. 662–671.

[23] K. Tomanek, F. Laws, U. Hahn, and H. Schütze, "On Proper Unit Selection in Active Learning: Co-Selection Effects for Named Entity Recognition," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, June 2009, pp. 9–17.

[24] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Active Learning for Biomedical Citation Screening," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 173–182.

[25] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[26] M. R. Anderberg, *Cluster analysis for applications*, 1973.

[27] R. Xu, D. Wunsch *et al.*, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[28] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[29] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek, "A framework for clustering uncertain data," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1976–1979, 2015.

[30] G. Hamerly, "Making k-means even faster." in *SDM*. SIAM, 2010, pp. 130–140.

[31] H.-P. Kriegel, P. Krger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[33] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data." in *SDM*. SIAM, 2003, pp. 47–58.

[34] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of text-mining systems for biology: overview of the second biocreative community challenge," *Genome Biology*, vol. 9, no. Suppl 2, p. S1, 2008.

[35] D. Campos, S. Matos, and J. L. Oliveira, "Gimli: open source and high-performance biomedical name recognition," *BMC Bioinformatics*, vol. 14, no. 1, p. 54, 2013.

[36] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: An evaluation," *Machine Learning*, vol. 68, no. 3, pp. 235–265, Oct. 2007.

[37] X. Han and J.-j. Kim, "Active learning for ontological event extraction," in *6th International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2014, pp. 45–51.