Journal of Biomedical Informatics 58 (2015) 11–18

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A study of active learning methods for named entity recognition in clinical text

Yukun Chen [a], Thomas A. Lasko [a], Qiaozhu Mei [c,d], Joshua C. Denny [a,b], Hua Xu [e,a,*]

[a] Department of Biomedical Informatics, Vanderbilt University, School of Medicine, Nashville, TN, USA
[b] Department of Medicine, Vanderbilt University, School of Medicine, Nashville, TN, USA
[c] School of Information, University of Michigan, Ann Arbor, MI, USA
[d] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
[e] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

ARTICLE INFO

Article history:
Received 28 April 2015
Revised 8 September 2015
Accepted 9 September 2015
Available online 15 September 2015

Keywords:
Active learning
Machine learning
Clinical natural language processing
Clinical named entity recognition

ABSTRACT

Objectives: Named entity recognition (NER), a sequential labeling task, is one of the fundamental tasks for building clinical natural language processing (NLP) systems. Machine learning (ML) based approaches can achieve good performance, but they often require large amounts of annotated samples, which are expensive to build due to the requirement of domain experts in annotation. Active learning (AL), a sample selection approach integrated with supervised ML, aims to minimize the annotation cost while maximizing the performance of ML-based models. In this study, our goal was to develop and evaluate both existing and new AL methods for a clinical NER task to identify concepts of medical problems, treatments, and lab tests from the clinical notes.

Methods: Using the annotated NER corpus from the 2010 i2b2/VA NLP challenge that contained 349 clinical documents with 20,423 unique sentences, we simulated AL experiments using a number of existing and novel algorithms in three different categories including uncertainty-based, diversity-based, and baseline sampling strategies. They were compared with the passive learning that uses random sampling. Learning curves that plot performance of the NER model against the estimated annotation cost (based on number of sentences or words in the training set) were generated to evaluate different active learning and the passive learning methods and the area under the learning curve (ALC) score was computed.

Results: Based on the learning curves of *F*-measure vs. number of sentences, uncertainty sampling algorithms outperformed all other methods in ALC. Most diversity-based methods also performed better than random sampling in ALC. To achieve an *F*-measure of 0.80, the best method based on uncertainty sampling could save 66% annotations in sentences, as compared to random sampling. For the learning curves of *F*-measure vs. number of words, uncertainty sampling methods again outperformed all other methods in ALC. To achieve 0.80 in *F*-measure, in comparison to random sampling, the best uncertainty based method saved 42% annotations in words. But the best diversity based method reduced only 7% annotation effort.

Conclusion: In the simulated setting, AL methods, particularly uncertainty-sampling based approaches, seemed to significantly save annotation cost for the clinical NER task. The actual benefit of active learning in clinical NER should be further evaluated in a real-time setting.

1. Introduction

Clinical notes in electronic medical records (EMR) contain much important patient information. Natural language processing (NLP) technologies offer a solution to convert free text data in EMR into structured representations, thus supporting studies in the clinical domain, such as disease phenotypes and patient cohort identification [1,2], decision support [3], and drug repurposing [4]. Identification of clinical concepts or clinical named entity recognition (NER) is an important task for building clinical NLP systems. For example, much work has been done to extract clinically important entities from clinical text, such as diseases, medications, procedures, and laboratory tests [5–7].

Some existing clinical NLP systems not only extract various types of clinical entities, but also map them to concepts in the controlled vocabularies such as the Unified Medical Language System (UMLS) [8], including cTAKES [9], MedLEE [10], MetaMap [11], and

* Corresponding author at: School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA. Tel.: +1 713 500 3924.
E-mail address: hua.xu@uth.tmc.edu (H. Xu).

http://dx.doi.org/10.1016/j.jbi.2015.09.010
1532-0464/

KnowledgeMap [12]. Many of these systems rely on symbolic NLP approaches and can be applied to various information extraction tasks. Recent studies have shown that machine learning (ML) based models, which are trained on annotated data sets, have the potential to achieve better performance in clinical NER tasks. Patrick and Li [13] developed a machine learning model to extract medication-related entities with a *F*-measure of 85.65% for the evaluation of exact match medication entry, which was optimal relative to other participants in the 2009 i2b2 NLP challenge. Both de Brujin et al. [14] and Jiang et al. [15] systematically investigated ML-based approaches for recognizing broader types of clinical entities and presented their promising results of 85.23% and 83.91% in *F*-measure, respectively, as the top two teams in the clinical concept extraction task in 2010 i2b2 NLP/VA challenge. Conditional random field (CRF) and support vector machine (SVM), which were the most widely used ML models in NER tasks, could build the most effective clinical concept extraction systems [6].

ML-based approaches, however, often require large amounts of annotated corpora, which are time-consuming to build due to the manual effort required for the task. In the clinical domain, some tasks require domain experts (e.g. physicians or nurses) to annotate text; thus the cost of annotation could be very high. In the general English domain, pool-based active learning strategies [16] have benefited many NLP tasks which require annotation from a large pool of unannotated data to construct the supervised ML model, such as word sense disambiguation [17], text classification [18], and information extraction [19]. In recent years, several studies have also applied active learning to text processing tasks in the clinical domain. Figueroa et al. [20] validated active learning algorithms as a way to reduce the size of training sets to yield expected performance in medical text classification tasks on five datasets. We also developed and evaluated active learning paradigm on multiple biomedical NLP tasks, such as assertion classification of concepts in clinical text [21], supervised word sense disambiguation in MEDLINE [22], and high-throughput phenotyping tasks for EMR data [23]. The common conclusion of these studies is that active learning could reduce annotation cost while improving the quality of the classification model, as compared to the passive learning approach (random sampling).

Different from the general classification tasks, NER is a sequence labeling task. Therefore, the goal of active learning for NER would be to select informative sequences (e.g. sentences) from the pool. Thus, different methods are needed to measure the informativeness of sequences. In the literature, some AL studies particularly focused on NER tasks and provided insightful information for approach design. An AL study by Kim et al. [24] presented a new AL paradigm for NER that considered both uncertainty of the classifier and the diversity of the corpus. For uncertainty sampling, they implemented *N*-best sequence entropy, which was based on the *N* most likely label sequences for the unlabeled samples; for diversity sampling, they considered three levels information, including NP chunk, Part-of-Speech tag, and the word itself, to compute the similarity between sentences. The combined performance was better than random sampling. However, their diversity-based method alone did not outperform random sampling. Settles and Craven [19] conducted a large-scale empirical study of AL for NER by evaluating seventeen methods in six corpora. They used random sampling and long sentence sampling methods as two baselines, and multiple active learning methods, including six uncertainty-sampling approaches, six query-by-committee methods, and other methods such as information density, fisher information, and expected gradient length. Most of the active learning algorithms performed better than baselines, indicating the promise of AL in NER. One limitation of these existing studies is that they are simulated studies that assumed that the annotation cost for each sentence was same. In reality, however, annotation cost could be different from one sentence to another. Informative sentences selected by active learning algorithms (e.g. uncertainty sampling) could require more annotation time just because they are longer sentences. There have been mixed results for doing cost-sensitive active learning in the literature to tackle realistic annotation costs [25–27]. Nevertheless, all the above studies of AL in NER were from the open domains and to the best of our knowledge, there is no AL on clinical NER tasks.

There are other techniques also aiming to efficiently build the NER models or reduce annotation effort in practice. Online learning is designed for the large-scale data training when computing resources are limited. Compared to batch learning that tends to induce an optimal model by training all the available labeled data, online learning would rather quickly generate a model on the basis of every single fresh random sample in the large data stream. Bottou and LeCun [28] showed that adequate online learning algorithms could asymptotically outperform any batch learning algorithms. Goldberg et al. [29] presented a Bayesian model called OASIS to efficiently train models on the stream data based on online learning and semi-supervised learning while filtering incoming unlabeled data by active learning technique for the large-scale machine learning tasks. On the other hand, pre-annotation or machine-assisted annotation is a strategy to speed up the manual annotation process by giving human annotator(s) the machine-annotated data beforehand. However, some studies found that the pre-annotation could introduce potential bias and their results are mixed. Fort and Sagot [30] found that pre-annotation gains the annotation speed, quality, and reliability for the part-of-speech corpus annotation tasks although biases did appear. Lingren et al. [31] presented a series of experiments to conclude that dictionary-based pre-annotation could reduce the annotation cost for clinical NER without introducing bias in the annotation process. South et al. [32] also evaluated the effects of machine-assisted annotation for de-identification of clinical text and found, however, that the pre-annotation did not improve the annotation quality and offered statistically significant time-saving, compared to the manual annotation from scratch.

In this study, we conducted simulated active learning experiments using an existing clinical NER corpus with annotated medical problems, treatments, and lab tests in clinical notes. We assessed six existing AL algorithms and developed seven novel AL algorithms for the clinical NER task. In addition to the traditional assumption of same annotation cost per sentence, we also evaluated our methods based on the assumption of same annotation cost per word, which is closer to the real world scenario. The results of our study showed that multiple active learning algorithms outperformed passive learning in both ways of evaluation.

2. Methods

2.1. Dataset

In this study, we used the annotated training corpus from the 2010 i2b2/VA NLP challenge, which contains 349 clinical documents with 20,423 unique sentences. Three types of medical entities: problem, treatment, and test, were annotated in each sentence. Table 1 shows the descriptive statistics of the corpus. The dataset is divided into two pieces: (1) the pool of data to be queried and (2) the independent test set for evaluation. As we used 5-fold cross validation in the experiment, the pool contains 80% of data randomly selected from the original dataset while the independent test set has the remaining 20% of the sentences.

2.2. Machine learning-based NER

To run the machine learning-based NER, we converted the annotations to the "BIO" format, where "B" represents the label

Table 1
Distribution of words and different types of entities and entity words in the corpus of 20,423 unique sentences.

	Overall count	Mean of count per sentence	SD of count per sentence
Word	225,670	11.05	9.73
Entity	26,206	1.28	1.65
Problem entity	11,192	0.55	1.03
Treatment entity	8099	0.40	0.91
Test entity	6915	0.34	1.02

for the beginning of an entity, "I" for inside the entity, and "O" for outside the entity. As we have three types of entities in our task, there are seven individual labels for identification, such as "B-problem", "B-treatment", "B-test", "I-problem", "I-treatment", "I-test", and "O".

In our previous study, we developed an ML-based NER system for the i2b2 dataset, with optimized features and ML algorithm [15]. For this study, we used the same set of optimized features and the same conditional random field (CRF) [33] classifier that is implemented in the CRF++ package [34].

2.3. Active learning experimental framework

In this study, we simulated the practical pool-based AL framework. Although all sentences in our corpus were pre-annotated, we did not utilize their labels unless the querying algorithms selected them. The following is the framework we used in the experiments:

(1) *Initial model generation:* At the beginning, a small number of samples are queried for annotation to build the initial model. We conducted experiments to compare two initial sampling strategies: (a) random sampling, and (b) longest sentence sampling. We decided on the strategy of using longest sentences to generate initial models because it could induce a better initial model or a starting point in the learning curve than random sampling for each querying method. Nevertheless, the conclusions based on strategy (a) or (b) stay the same as long as all different methods used the same set of initial samples for a fair comparison.
(2) *Querying:* The unannotated sentences were then ranked based on the querying algorithm. Some algorithms require the updated CRF model for ranking (e.g. uncertainty sampling) while some do not (e.g. all diversity based algorithms). The top ranked sentences were selected for annotation, and then added to the annotated set. In our experiment, the batch sizes (the size of top ranked sentences selected for annotation) of each iteration were $8,16,32,64,128,\ldots,2^{(i+2)}$, where i is the number of iterations. This is one of the standard ways to select batch size for the active learning experiment and has been used in an active learning challenge [35].
(3) *Training:* The CRF model was retrained on the updated annotated set.
(4) *Iteration:* Steps (2) and (3) were repeated until the stop criterion was met. In this study, the annotation stops when all sentences in the pool of unlabeled set were queried.

Multiple measurements were stored during the active learning process for evaluation, such as model quality in *F*-measure, number of words in the annotated set, and number of entities in the annotated set.

As shown above, querying algorithms are critical for an active learning system. The following sections discuss three types of querying algorithms that we investigated in our experiments. Some algorithms were developed by previous studies and some algorithms are newly developed in this study (marked as new).

2.3.1. Uncertainty-based querying algorithms

The assumption here is that the most uncertain sentences are most informative because identification of their uncertain labels could gain the most utility for the supervised NER learning. We considered a label of a sentence as a sequence of labels of words. In most of our implementations, only the *N*-best sequence labels were considered since the size of the possible sequence labels grows exponentially as the length of a sentence increases. We also extended the *N*-best sequence labels to cover most of the highly possible labels. The entropy of words and entities were also tested in our study. The six methods we implemented to calculate the uncertainty of a sentence are described below:

(1) Least Confidence (LC): to take the uncertainty from the best possible sequence label based on the posterior probability output from CRF. The uncertainty of a sentence x is equal to $1 - P(y*|x)$, where y^* is the most likely sequence label for the sentence x.
(2) Margin: to take the uncertainty from the best two possible sequence labels. The uncertainty of a sentence x is equal to $P(y^*|x) - P(y^{**}|x)$, where y^* and y^{**} are the most likely and second most likely sequence labels, respectively, for the sentence x. The lower margin between the two probabilities represents higher uncertainty.
(3) N-best sequence entropy: to take the entropy of the probability distribution over *N*-best sequence labels predicted by the CRF model. We used $N = 3$ in our experiments.
(4) Dynamic N-best sequence entropy (new): to take the *N*-best sequence labels with the sum of their probabilities being at least 0.9. Here, *N* ranges from 1 to 20 in our experiments. For example, if the best sequence label has a probability of 0.95, *N* is equal to 1 (equivalent to LC); if the best 4 sequence labels have probabilities of 0.4, 0.3, 0.1, and 0.1, the sum of the probabilities is 0.9 and therefore, *N* is 4 and we ignore the 5th and later labels.
(5) Word entropy: to take the summation of entropy of all individual words in a sentence: $WordEntropy(x) = \sum_{i=1}^{n} Entropy(w_i)$, where w_i is the ith word in the sentence x with n words. The entropy of an individual word is calculated based on the distribution of seven possible labels for an individual word: $Entropy(w_i) = -\sum_{j=1}^{7} P(y_j|w_i) log\, P(y_j|w_i)$, where the seven possible labels (y_1–y_7) are listed in Section 2.2.
(6) Entity entropy (new): to take the summation of entropy of only the beginning words of the estimated entities (e.g. B-entity). This is a simple heuristic method that builds on "*Word entropy*". Instead of considering that all the words in a sentence equally contribute to NER model building, we think entropies of entity words are probably more important. Moreover, we believe that entropy of "B" words are more important then "I" words. Therefore, the "*Entity Entropy*" only considers the entropies from the words with B-entity estimated by the NER model.

2.3.2. Diversity-based querying algorithms

Uncertainty sampling is highly dependent on the quality of the model. Therefore, it may not be efficient in a practical setting where updating the model may take time. In this section, we propose diversity-based querying algorithms that consider the

information other than the model, such as the similarity between sentences.

The idea behind the diversity-based querying algorithms is that we do not want to query the sentences that are similar to those that are already annotated. We applied the vector space model to pre-calculate pair-wise cosine similarity of any two sentences in the corpus. We used complete-linkage (max similarity) to determine the similarity between an unlabeled sentence and a group of labeled sentences. Unlabeled sentences with lower similarity scores would be assigned higher priority to be selected for annotation. The advantages of the diversity-based algorithms are (1) it is not dependent on the model and the annotation results; (2) the pair-wise similarity scores between sentences could be pre-computed, thus the querying step could be very efficient.

To find the best similarity measurements, we explored different features at the word, semantic, and syntactic levels for building vectors and calculating similarity scores. We also combined all of them for better similarity assessment.

(1) *Word similarity* (new): A vector of words weighted by the TF/IDF weighting scheme is used to represent each sentence. Then the cosine similarity between two vectors is calculated as the similarity between the two sentences.

(2) *Syntax similarity* (new): Each sentence is parsed by the Stanford parser [36] and the dependency relations derived from the parse tree are used to form the vector. For example, a sentence "She is afebrile with stable vital signs." has six dependencies "nsubj(afebrile-3, She-1)", "cop(afebrile-3, is-2)", "prep(afebrile-3, with-4)", "amod(signs-7, stable-5)", "amod(signs-7, vital-6)", and "pobj(with-4, signs-7)". To generalize the dependency relations, we then replaced the arguments of relations by their corresponding part of speech (POS) tags. The above example was converted into a vector of ["nsubj(JJ, PRP)", "cop(JJ, VBZ)", "prep(JJ, IN)", "amod(NNS, JJ)", "amod(NNS, JJ)", and "pobj(IN, NNS)"]. We weighted each dependency relation in the vector using the TF/IDF weight scheme based on their counts in the sentence and the corpus. Finally, cosine similarity was computed for each pair of sentences, similar to the method of word similarity.

(3) *Semantic similarity* (new): This method is to calculate semantic similarity between two sentences based on concept similarity. We modified an existing semantic similarity method originally based on word similarity [37]. Our approach consisted of two steps: (1) extraction of clinical concepts in each sentence so that each sentence can be represented using a vector of union concepts from the two sentences; and (2) calculation of the similarity between the two sentence vectors of concepts, by measuring similarity scores between any two concepts and computing the cosine similarity of two sentence vectors. For Step 1, we processed each sentence using KnowledgeMap Concept Identifier (KMCI) [38], a general clinical NLP system, that extracts clinical concepts defined in the UMLS. Each sentence was represented by a vector of UMLS concept unique identifiers (CUIs) of the union concepts. For Step 2, the semantic similarity (or distance) between any two UMLS concepts was calculated using the package of *UMLS-interface* and *UMLS-similarity* [39], which computes the similarity between two CUIs by using the user-selected similarity measurement (i.e. *Path*, *LCH* [40], *WUP* [41], etc.) with a specified source (i.e. SNOMED-CT [42] and MeSH [43]). The value of each union concept of a sentence is the max similarity among the similarity scores between each of the concepts from this sentence and this union concept. Once we formed the semantic vector for two sentences, we computed the cosine similarity between them. Fig. 1 demonstrates an example of how *semantic similarity* is calculated for two sentences: S1: "You will need to have your uterine bleeding evaluated." and S2: "This continued agitation may be caused by intraparenchymal hemorrhage."

KMCI identified "uterine bleeding" as a UMLS concept (with CUI: C0042134) in S1 and "continued agitation" (C0085631) and "intraparenchymal hemorrhage" (C0019080) in S2. The union UMLS concepts of the two sentences are C0085631, C0019080, and C0042134. Then we applied UMLS-similarity package to compute the similarity between each of the two concepts. The vector for S1 is [0.13,0.5,1], where 0.13 is the UMLS similarity between C0042134 and C0085631, 0.5 is for the one between C0042134 and C0019080, and 1 is for the one between C0042134 and C0042134. The vector for S2 is [1,1,0.5], where 0.5 is the UMLS similarity between C0042134 and C0019080, the first 1 is for the one between C0085631 and C0085631, and the second 1 is for the one between C0019080 and C0019080. Then the similarity between S1 and S2 is 0.67, which is the cosine similarity of these two vectors.

(4) *Combined similarity* (new): This approach combines all word, syntactic, and semantic information for similarity calculation. We first combined words and dependency relations for the same sentence into one vector, and then computed the cosine similarity for each pair of sentences based on the new vectors. The final combined similarity between the two sentences is the average similarity for both the newly computed cosine similarity between word/dependency vectors and the semantic similarity based on UMLS.

In principle, zero similarity score would indicate very diverse sentences, which we want to select. However, after careful analysis, we found that sentences with a zero similarity score to the labeled set were usually short sentences, which contain very few clinical entities. For example, short sentences such as section headers contain few dependency relations and often yield zero syntax similarity. Therefore, we decided to eliminate unlabeled sentences with zero similarity to the labeled set from sample selections for all diversity-based algorithms.

2.3.3. Baseline algorithms

In addition, we also included two querying algorithms that simply consider the length of the words or entities in a sentence. As we mentioned in the introduction, one limitation of such simulated active learning studies is to assume that each sentence costs the same amount of annotation effort, which obviously is not true in reality. By including these two extremely biased methods as additional baselines, we hope to further confirm the effectiveness of AL methods.

(1) *Length-words* is a simple querying method that selects sentences with the largest number of words. The assumption is simply that longer sentences may contain more information for NER than shorter ones.

(2) *Length-concepts* is another simple querying method that selects sentences with the largest number of clinical concepts, as identified by KMCI. The assumption is that sentences with more clinical concepts are more informative sentences for NER.

In addition, we included the typical passive learning method *Random*, which randomly selects samples at each iteration.

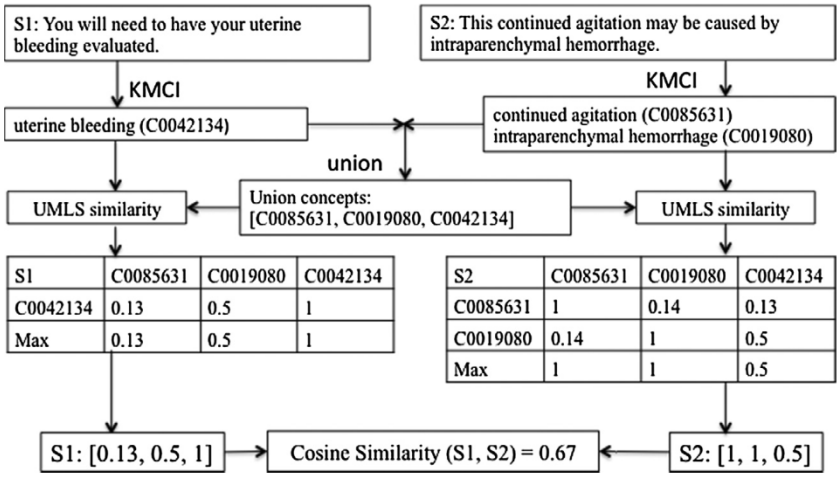

Fig. 1. An example of computing similarity between two sentences using *semantic similarity* algorithm.

2.4. Evaluation

Most of the active learning studies utilized learning curves that plot *F*-measure of the model on an independent test set as a function of sample size of the training set as the primary evaluation approach. Following previous studies on open domain NER [19,24], we first evaluated our AL-enabled clinical NER using the same type of learning curve that plots *F*-measure versus number of annotated sentences, assuming annotation cost is same for each sentence. However, we think the annotation costs for different sentences could be greatly different in reality; thus simply assuming the equal annotation cost of each sentence, as the traditional way does, could induce an inaccurate estimation about the benefit of active learning in reality. Therefore, we also generated the learning curve of *F*-measure versus number of words in the annotated sentences as a new assessment approach. The new way of word-based evaluation assumes that the annotation cost is proportional to the length of sentence, and is therefore a better way to estimate the real annotation cost. We also computed the area under the learning curve (ALC) as a global score for both evaluation methods, which was a major metric to evaluate active learning methods in the challenge [35]. The ALC scores for the learning curves of *F*-measure vs. sentences and *F*-measure vs. words are labeled as ALC1 and ALC2, respectively. To further demonstrate some characteristics of different querying methods, we plotted additional curves, including the *entity count curve* that plots the number of entities versus the number of sentences and the *sentence length curve* that plots the number of words (length of sentences) versus the number of annotated sentences.

Our evaluation results were based on 5-fold cross validation (CV). For each iterative experiment, one fold was used as an independent test set and four other folds were used as the pool of querying and training set. The results on the learning curves were averaged over the five runs. For the experiments using random sampling, we repeated the experiments of 5-fold CV five times and averaged their results.

3. Results

All methods were tested in the same active learning framework and cross validation setting (e.g. the same initial queries and model, pool, batch size, parameters of CRF model, and test set). Table 2 shows the ALC scores based on two types of learning curves for twelve active learning algorithms in three categories and *Random* that represents passive learning.

For ALC1 that is based on learning curves of *F*-measure vs. number of sentences, all active learning algorithms, except *syntax similarity*, were better than random sampling. Among the three types of algorithms, uncertainty-based sampling methods (0.83 in

Table 2
Two types of ALC scores for all active learning algorithms versus passive learning.

Categories	Methods	Existing or New	ALC1 score	ALC2 score
			F-measure vs. Sentences	*F*-measure vs. Words
Uncertainty based sampling methods	*LC*	Existing	0.83	0.84
	Margin	Existing	0.83	0.84
	N-best sequence entropy	Existing	0.81	0.85
	Dynamic N-best sequence entropy	New	0.82	0.84
	Word entropy	Existing	0.83	0.84
	Entity entropy	New	0.83	0.84
Diversity based sampling methods	*Word similarity*	New	0.77	0.82
	Syntax similarity	New	0.72	0.80
	Semantic similarity	New	0.79	0.83
	Combined similarity	New	0.76	0.82
Baseline methods	*Length-Words*	Existing	0.82	0.81
	Length-Concepts	New	0.82	0.81
Passive Learning	*Random*	Existing	0.74	0.82

Note: ALC1 is the ALC (area under the learning curve) score for the learning curves of *F*-measure vs. number of sentences; ALC2 is the ALC score for the learning curves of *F*-measure vs. number of words.

average ALC1) outperformed two baselines (0.82 in average ALC1), which outperformed diversity-based methods (0.76 in average ALC1).

For ALC2 that is based on learning curves of *F*-measure vs. number of words, three types of querying algorithms performed differently: all six uncertainty-based methods outperformed random sampling; in the diversity sampling category, only *semantic similarity* achieved better performance than *Random*; ALC2 of baseline methods (*Length-Words* and *Length-Concepts*) did not exceed random sampling because the tendency of selecting longest sentences was penalized in this evaluation.

We generated two types of learning curves for all thirteen methods. However, it is too crowded to plot thirteen curves in one diagram. Therefore, we selected best-performing method in each category to display their learning curves versus *Random*. Fig. 2 shows the traditional learning curves based on *F*-measure versus number of annotated sentences for methods of *LC*, *semantic similarity*, *Length-concepts*, and *Random*. The method of *Length-concepts* had the best performance at the very early stage of learning curves, but was surpassed by *LC* at the later stages, which outperformed the other methods. Fig. 3 shows the new type of learning curves based on *F*-measure versus number of words in the annotated sentences for the methods of *N-best sequence entropy*, *semantic similarity*, *Length-concepts*, and *Random*. *N-best sequence entropy* led all the stages of active learning.

Based on learning curves in Figs. 2 and 3, we also computed the number of annotated sentences and words required to achieve a fixed *F*-measure for each method. We used linear interpolation to estimate the points from learning curves that were not actually available. Furthermore, we estimated the extent of annotation cost saving achieved by active learning as compared to passive learning for achieving the same performance. For example, to achieve 0.80 in *F*-measure, *LC* used 2971 sentences or 61,238 words, *N-best sequence entropy* required 3249 sentences or 62,486 words, *semantic similarity* needed 5468 sentences or 98,075 words, *Length-concepts* required 5201 sentences or 109,580 words, and *Random* queried 8702 sentences or 105,340 words. Compared to *Random* with respect to cost saving in sentences, *LC* saved 5731 sentences (66%), *semantic similarity* saved 3234 sentences (37%), and *Length-concepts* saved 3501 sentences (40%). With respect to cost saving in words, *N-best sequence entropy* reduced 42,854 words (41% saving of annotation cost in words), *LC* could save more – 44,102 words (42%). However, *semantic similarity* saved only 7265 words (7%), and *Length-concepts* actually required annotating 4240 additional words (4% increase of annotation cost in words).

In addition to learning curves, we also show two characteristics of methods to measure how informative the queried sentences are.

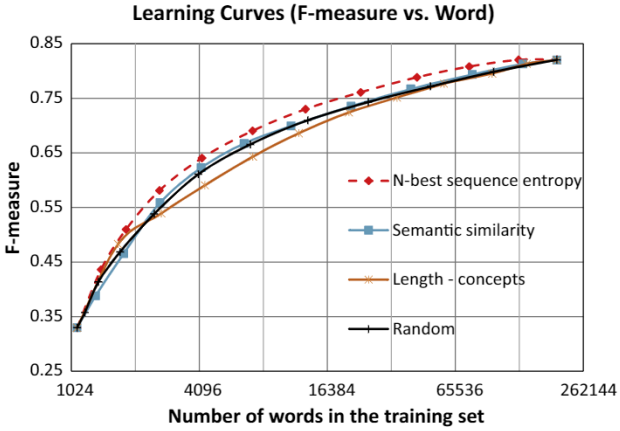

Fig. 3. Learning curves for *F*-measure versus Word.

The *entity count curve* reports the total *entity count* at each iteration of the active learning process. The *sentence length curve* reports the total number of words at each iteration. We could globally measure the characteristics of *entity count* and *sentence length* for each method based on the area under the *entity count curve* and area under the *sentence length curve*, respectively. Fig. 4 shows the entity count curves for *Random* and other methods (*LC*, *semantic similarity*, *Length-concepts*) that achieved the largest area under the *entity count curve* in their categories. Fig. 5 shows the sentence length curves for *Random* and other methods (*LC*, *semantic similarity*, *Length-words*) that achieved the largest area under the *sentence length* curve in their categories.

Both *entity count curves* and *sentence length curves* present a very similar pattern of the methods. *Length-concepts* and *Length-words* queried the most number of concepts per sentence and the longest sentences, respectively, at all stage of active learning, while random sampling did the least. Both *LC* and *semantic similarity* are in between the curves mentioned above, but they both performed better than *Length-concepts*, *Length-words*, and *Random* in terms of ALC2.

4. Discussion

In this study, we conducted simulated active learning experiments for a clinical NER task and demonstrated that active learning has the potential to reduce annotation cost for building clinical NER models. To the best of our knowledge, this is the one of the

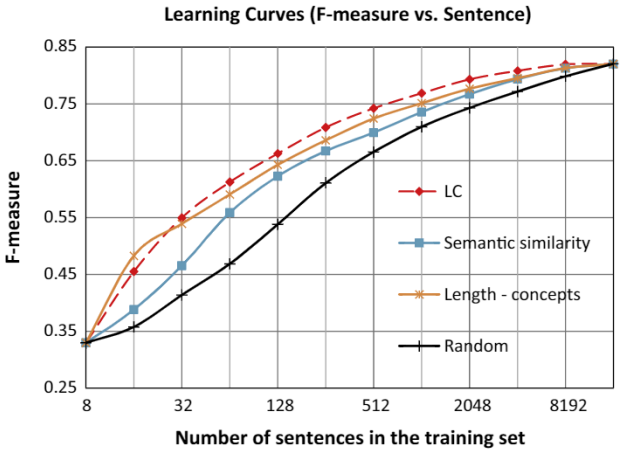

Fig. 2. Learning curves for *F*-measure versus Sentence.

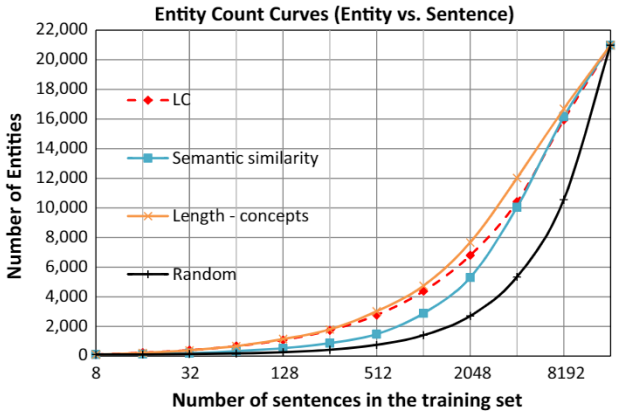

Fig. 4. Entity count curves that plot number of entities versus number of sentences in the training set.

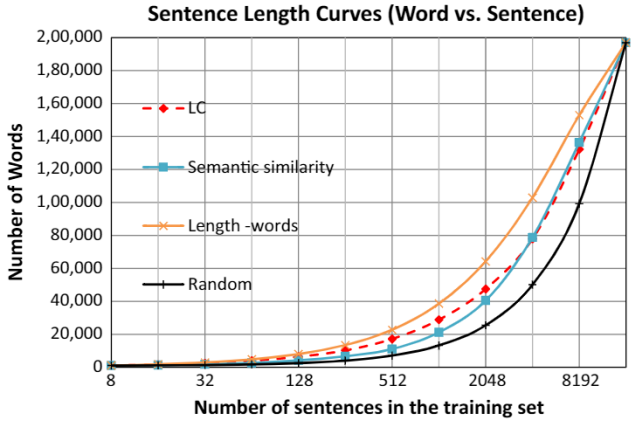

Fig. 5. Sentence length curves that plot number of words versus number of sentences in the training set.

earliest studies on active learning for clinical NER. According to Figs. 2–5, active learning algorithms (e.g. uncertainty sampling) did query longer sentences with higher number of entities per sentence, which could contribute to higher ALC1 and ALC2 scores. However, simply selecting sentences with high number of entities (e.g. *Length-concepts*) or longest sentences (e.g. *Length-words*) failed to surpass passive learning in ALC2 score, which we consider as a closer evaluation metric in the real-time situation. This finding suggests that active leaning does select informative samples that can help build better clinical NER models quickly.

Uncertainty-sampling based algorithms outperformed all other methods in both ALC1 and ALC2, because they queried the most informative sentences using the knowledge of trained models. Among these methods, *LC*, *Margin*, *Word Entropy*, and *Entity Entropy* had very similar results (0.83 in ALC1 and 0.84 in ALC2). *N-best sequence entropy* gained highest ALC2 (0.85), indicating that it is probably more efficient in reality. However, one concern of applying uncertainty sampling based methods to real-world annotation tasks is that they rely on the updated NER models, which may take time when the annotated data are getting bigger. For example, it would take several minutes to fully train a model based on 1000 annotated sentences in our experiment. In reality, it may not be feasible to ask annotators to wait such a long time for the next iteration of queried samples.

The diversity sampling methods, on the other hand, do not depend on the CRF model and most processes can be precomputed before the annotation process starts, which makes it more appealing. However, the current diversity-based methods implemented in this study did not perform as well as the uncertainty sampling. Another research direction is to combine uncertainty and diversity methods, e.g. using the linear function from Kim et al. [24] or nonlinear function from Settles and Craven [19]. However, we explored both above combination methods and they did not achieve better performance than the uncertainty method alone on our dataset. Another combination approach is to integrate clustering algorithms (e.g. *k-means* or *affinity propagation* [44]) with uncertainty sampling to find the most uncertain and representative samples.

The active learning algorithms developed and evaluated in this study should work for other supervised NER models, such as Maximum Entropy (ME) [45], support vector machine (SVM) [46], and structural SVM (SSVM) [47]. They were also studied in the clinical NER tasks [15,48,49] in addition to the CRF model. For uncertainty sampling, the required inputs are the probabilities of the estimated sequence of labels or the probabilities of the estimated label for each individual word, which could be derived from ME, SVM, or SSVM. For diversity sampling, the statistical language models are not even necessary.

Another contribution of this work is to introduce a new evaluation metric for simulated active learning studies for NER. Instead of assuming that each sentence requires the same amount of annotation effort, we assume each word requires the same amount of annotation effort. Therefore, the estimated savings of annotation cost in our study would be closer to the reality, where longer sentences probably need more annotation time than the shorter ones. Our results seem to support this intuition. For example, to achieve an *F*-measure of 80%, the *LC* method could save 66% sentences; but the saving would be only 42% if we consider words instead of sentences. The 24% drop of savings indicates that the traditional way of evaluation could overestimate the effectiveness of active learning methods in NER, when compared to passive learning. Moreover, other active learning methods such as the diversity sampling methods, which could outperform passive learning in ALC1, did not achieve the same performance when ALC2 was used in evaluation. For example, the *semantic similarity* method showed a saving of 37% in ALC1 evaluation; but it had a saving of only 7% in ALC2 evaluation. These findings suggest that we should be more cautious about results from simulated experiments of active learning on clinical NER. The actual benefit of active learning should be further evaluated using real-time settings of NER tasks.

As described above, the main limitation of this study is that it is a simulated study of active learning for clinical NER. To assess the real value of active learning for clinical NLP, we will have to evaluate it in a real-world setting. There are a few machine learning systems with integrated active learning components, such as the DUALIST system [50] for word sense disambiguation in open domains. However, to our knowledge, there is no clinical NLP system that integrates a practical active learning module. Therefore, our next step is to develop a clinical NER system, which consists of an annotation interface and an active learning component that actively selects samples for annotation. We will then conduct formal user studies to compare active learning vs. passive learning in terms of annotation time and model quality.

5. Conclusion

We conducted a simulated study to compare different active learning algorithms for a clinical NER task. Our results showed that most active learning algorithms outperformed the passive learning method when we assume equal annotation cost for each sentence. However, savings of annotation by active learning were reduced when the length of sentences was considered. We suggest that the effectiveness of active learning for clinical NER needs to be further evaluated by developing active learning enabled annotation systems and conducting user studies.

Conflict of interest

None.

Acknowledgment

This study was supported by the National Institutes of Health in United States (NIH) Grant NLM 2R01LM010681-05.

References

[1] O. Gottesman, H. Kuivaniemi, G. Tromp, W.A. Faucett, R. Li, T.A. Manolio, et al., The electronic medical records and genomics (eMERGE) network: past, present, and future, Genet. Med. 15 (2013) 761–771.
[2] H. Xu, Z. Fu, A. Shah, Y. Chen, N.B. Peterson, Q. Chen, et al., Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases, AMIA Annu. Symp. Proc. 2011 (2011) 1564–1572.

[3] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can natural language processing do for clinical decision support?, J Biomed. Inform. 42 (2009) 760–772.
[4] H. Xu, M.C. Aldrich, Q. Chen, H. Liu, N.B. Peterson, Q. Dai, et al., Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality, J. Am. Med. Inform. Assoc. 22 (2015) 179–191.
[5] O. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (2010) 514–518.
[6] O. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (2011) 552–556.
[7] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, J. Am. Med. Inform. Assoc. 20 (2013) 806–813.
[8] NIH, Unified Medical Language System (UMLS). <http://www.nlm.nih.gov/research/umls/>.
[9] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[10] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson, P.D. Clayton, Unlocking clinical data from narrative reports: a study of natural language processing, Ann. Intern. Med. 122 (1995) 681–688.
[11] A.R. Aronson, F.M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (2010) 229–236.
[12] J.C. Denny, R.A. Miller, K.B. Johnson, A. Spickard 3rd, Development and evaluation of a clinical note section header terminology, AMIA Annu. Symp. Proc. (2008) 156–160.
[13] J. Patrick, M. Li, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, J. Am. Med. Inform. Assoc. 17 (2010) 524–527.
[14] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, X. Zhu, Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010, J. Am. Med. Inform. Assoc. 18 (2011) 557–562.
[15] M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, et al., A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Inform. Assoc. 18 (2011) 601–606.
[16] D.D. Lewis, W.A. Gale, A sequential algorithm for training text classifiers, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 3–12.
[17] J. Zhu, E. Hovy, Active learning for word sense disambiguation with methods for addressing the class imbalance problem, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 783–790.
[18] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, J. Mach. Learn. Res. 2 (2002) 45–66.
[19] B. Settles, M. Craven, An analysis of active learning strategies for sequence labeling tasks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008, pp. 1069–1078.
[20] R.L. Figueroa, Q. Zeng-Treitler, L.H. Ngo, S. Goryachev, E.P. Wiechmann, Active learning for clinical text classification: is it better than random sampling?, J Am. Med. Inform. Assoc. 19 (2012) 809–816.
[21] Y. Chen, S. Mani, H. Xu, Applying active learning to assertion classification of concepts in clinical text, J. Biomed. Inform. 45 (2012) 265–272.
[22] Y. Chen, H. Cao, Q. Mei, K. Zheng, H. Xu, Applying active learning to supervised word sense disambiguation in MEDLINE, J. Am. Med. Inform. Assoc. 20 (2013) 1001–1006.
[23] Y. Chen, R.J. Carroll, E.R. Hinz, A. Shah, A.E. Eyler, J.C. Denny, et al., Applying active learning to high-throughput phenotyping algorithms for electronic health records data, J. Am. Med. Inform. Assoc. 20 (2013) e253–e259.
[24] S. Kim, Y. Song, K. Kim, J.-W. Cha, G.G. Lee, MMR-based active machine learning for bio named entity recognition, in: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, Association for Computational Linguistics, New York, New York, 2006, pp. 69–72.
[25] A. Kapoor, E. Horvitz, S. Basu, Selective supervision: guiding supervised learning with decision-theoretic active learning, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., Hyderabad, India, 2007, pp. 877–882.
[26] S. Arora, E. Nyberg, C.P. Rosé, Estimating annotation cost for active learning in a multi-annotator environment, in: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 18–26.
[27] R. Haertel, E. Ringger, K. Seppi, J. Carroll, P. McClanahan, Assessing the costs of sampling methods in active learning for annotation, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 65–68.
[28] L. Bottou, Y. LeCun, Large scale online learning, in: L.S.a.B.S. Sebastian Thrun (Ed.), Advances in Neural Information Processing Systems, vol. 16, MIT Press, 2004.
[29] A.B. Goldberg, X. Zhu, A. Furger, J.-M. Xu, OASIS: Online Active Semi-Supervised Learning, 2011.
[30] K. Fort, B. Sagot, Influence of pre-annotation on POS-tagged corpus development, in: Proceedings of the Fourth Linguistic Annotation Workshop, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 56–63.
[31] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J. Meinzen-Derr, M. Kaiser, et al., Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements, J. Am. Med. Inform. Assoc. 21 (2014) 406–413.
[32] B.R. South, D. Mowery, Y. Suo, J. Leng, O. Ferrandez, S.M. Meystre, et al., Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text, J. Biomed. Inform. 50 (2014) 162–172.
[33] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. of the 18th International Conf. on Machine Learning, Morgan Kaufmann, Williamstown, MA, 2001, pp. 282–289.
[34] http://crfpp.googlecode.com/svn/trunk/doc/index.html.
[35] Active Learning Challenge, 2010. <http://www.causality.inf.ethz.ch/activelearning.php>.
[36] R. Socher, J. Bauer, C.D. Manning, A.Y. Ng, Parsing with Compositional Vector Grammars, ACL, 2013.
[37] Y. Li, D. McLean, Z.A. Bandar, J.D. O'Shea, K. Crockett, Sentence similarity based on semantic nets and corpus statistics, IEEE Trans. Knowl. Data Eng. 18 (2006) 1138–1150.
[38] J.C. Denny, J.D. Smithers, R.A. Miller, A. Spickard 3rd., "Understanding" medical school curriculum content using KnowledgeMap, J. Am. Med. Inform. Assoc. 10 (2003) 351–362.
[39] B.T. McInnes, T. Pedersen, S.V. Pakhomov, UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: AMIA Annu. Symp. Proc., vol. 2009, 2009, pp. 431–435.
[40] C. Leacock, G.A. Miller, M. Chodorow, Using corpus statistics and WordNet relations for sense identification, Comput. Linguist. 24 (1998) 147–165.
[41] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133–138.
[42] NIH, SNOMED Clinical Terms (SNOMED CT). <http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html>.
[43] NIH, MeSH. <http://www.nlm.nih.gov/mesh/meshhome.html>.
[44] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[45] A.E. Borthwick, A Maximum Entropy Approach to Named Entity Recognition, New York University, 1999.
[46] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[47] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (2005) 1453–1484.
[48] B. Tang, H. Cao, Y. Wu, M. Jiang, H. Xu, Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features, BMC Med. Inform. Decis. Mak. (2013).
[49] S. Doan, H. Xu, Recognizing medication related entities in hospital discharge summaries using support vector machine, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Beijing, China, 2010, pp. 259–266.
[50] B. Settles, Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, United Kingdom, 2011, pp. 1467–1478.