Curatable Named-Entity Recognition Using Semantic Relations

Yi-Yu Hsu and Hung-Yu Kao

Abstract—Named-entity recognition (NER) plays an important role in the development of biomedical databases. However, the existing NER tools produce multifarious named-entities which may result in both curatable and non-curatable markers. To facilitate biocuration with a straightforward approach, classifying curatable named-entities is helpful with regard to accelerating the biocuration workflow. Co-occurrence Interaction Nexus with Named-entity Recognition (CoINNER) is a web-based tool that allows users to identify genes, chemicals, diseases, and action term mentions in the Comparative Toxicogenomic Database (CTD). To further discover interactions, CoINNER uses multiple advanced algorithms to recognize the mentions in the BioCreative IV CTD Track. CoINNER is developed based on a prototype system that annotated gene, chemical, and disease mentions in PubMed abstracts at BioCreative 2012 Track I (literature triage). We extended our previous system in developing CoINNER. The pre-tagging results of CoINNER were developed based on the state-of-the-art named entity recognition tools in BioCreative III. Next, a method based on conditional random fields (CRFs) is proposed to predict chemical and disease mentions in the articles. Finally, action term mentions were collected by latent Dirichlet allocation (LDA). At the BioCreative IV CTD Track, the best F-measures reached for gene/protein, chemical/drug and disease NER were 54 percent while CoINNER achieved a 61.5 percent F-measure. System URL: http://ikmbio.csie.ncku.edu.tw/coinner/introduction.htm.

Index Terms—Biomedical text mining, curated term identification, named-entity recognition

1 Introduction

Named-entity recognition (NER) has become increasingly important in the area of biocuration, especially in the biomedical field. Biologists have applied many controlled vocabularies to annotate features from genotype to phenotype [1], [2]. A single data source is insufficient for predicting oncogenes accurately while multiple data sources provide more accurate predictions. Hence, the NER task is a prerequisite for the analysis of high-throughput screen and cross-referencing in different databases and heterogeneous data repositories [3], [4]. Several studies have shown that biomedical ontologies can be used to significantly improve literature searching on multiple domains [5], [6]. For example, medical subject heading (MeSH) is a well-known ontology maintaining good quality of literature searching in PubMed. PubMed was developed by the National Library of Medicine and contains a huge amount of literature, experimental results, and ontology information. The purpose of MeSH is to index journal articles and improve search results in PubMed, from which researchers gain a clearer understanding of biological processes and human health.

Most researchers have focused on the issue of NER-related tasks separately, and only a few works have considered the issue of integrating multiple NER. Therefore, we constructed an entity co-occurrence analysis system, which is suitable for gene, chemical and disease queries. Biomedical terms are often correlated with their co-occurrence in a sentence. This co-occurrence indicates that these biomedical terms both occur in the same paper. Although co-occurrence does not guarantee that the biomedical terms are relevant, co-occurrence features are still useful for providing possible candidates for relation extraction. Interest in assisting manual curation has dramatically increased [7]. The Critical Assessment of Information Extraction Systems in Biology (BioCreative) 2012 workshop initiated literature triage, in which participants designed systems to effectively prioritize articles for curation [8]. To further analyze manual curation, BioCreative IV focused on the integration of text mining and biology communities for the purpose of developing practically relevant biocuration systems. In the BioCreative IV workshop, the organizers hosted a web services-based NER track to identify gene/protein, chemical/drug, disease and action term mentions, which facilitates biocuration in PubMed abstracts [9].

When a manual curation database, such as the Comparative Toxicogenomics Database (CTD) [10], was developed, biocurators used PubMed and MeSH to annotate biomedical terms. Beyond effectively finding the articles required by biocurators, NER has therefore become an important task. For example, CTD developers organized document ranking and NER-related tasks for the CTD text-mining pipeline in a BioCreative competition. The CTD staff used a set of third party NER tools: Abner [11] for gene NER, Oscar3 [12] for chemical NER, and MetaMap [13] for disease recognition, as well as medical terms related to chemicals and genes. These NER tools are important to the development of CTD's controlled vocabulary terms in the text of an article's abstract. Thus, CTD-related NER tasks are critical

- *The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C. E-mail: Alan.Hsu@ikmlab.csie.ncku.edu.tw, hykao@mail.ncku.edu.tw.*

Manuscript received 5 Oct. 2014; revised 14 Oct. 2014; accepted 28 Oct. 2014. Date of publication 2 Nov. 2014; date of current version 4 Aug. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2366770

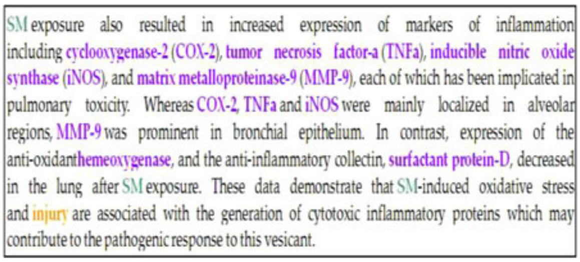

SM exposure also resulted in increased expression of markers of inflammation including cyclooxygenase-2 (COX-2), tumor necrosis factor-a (TNFa), inducible nitric oxide synthase (iNOS), and matrix metalloproteinase-9 (MMP-9), each of which has been implicated in pulmonary toxicity. Whereas COX-2, TNFa and iNOS were mainly localized in alveolar regions, MMP-9 was prominent in bronchial epithelium. In contrast, expression of the anti-oxidanthemeoxygenase, and the anti-inflammatory collectin, surfactant protein-D, decreased in the lung after SM exposure. These data demonstrate that SM-induced oxidative stress and injury are associated with the generation of cytotoxic inflammatory proteins which may contribute to the pathogenic response to this vesicant.

Fig. 1. A paragraph containing curatable and non-curatable named-entities.

for building annotated multiple named entities. However, the workflow of CTD-related NER is a heuristic learning procedure, and the heuristic rules are generated by the experience of the biocurators [14]. The biocurators annotate the biomedical terms based on their domain knowledge. Recently, PubTator was developed for accelerating and assisting manual literature curation [15]. PubTator offers biocurators for annotating research targets with essential NER results. Although PubTator emphasizes on an abundance of NER annotations, biocurators may be confused by redundant information, especially when investigating a CTD-related NER task. To approach such CTD-related NER problems, the BioCreative IV committee examined the curated terms made by biocurators when they were determining whether biomedical terms should be curated. The CTD-related NER task is easily understood by biocurators, but the procedure is difficult for machine learning.

The NER problem leads biocurators into problems that include time-consuming processes and an insufficient quantity of curated articles. Biocurators can only annotate a limited number of articles, and the huge amount of biomedical literature creates barriers to the biocuration process [16]. Therefore, supplying accurate information to biocurators is critical. Curatable named entities are considered as interaction pairs and are annotated by biocurators. As shown in Fig. 1, the orange named-entity is a disease; the cyan named-entity is a chemical, and the purple colored one is a gene. SM refers to sulfur mustard. In this case, cyclooxygenase-2 is identified as a gene named-entity, but it is not annotated such that cyclooxygenase-2 can interact with SM. That is, cyclooxygenase-2 is a non-curatable named-entity in this paragraph. When we observed the curatable and non-curatable named-entities in the BioCreative IV CTD NER task, we collected statistics for further analysis. Thirty-seven percent of the false positive and negative cases of our NER results are matched named-entities, but they are non-curatable ones. That is, the information extraction can be improved by reducing non-curatable named-entities. In this work, we aim to design a curatable sentence classifier that is capable of retrieving sentences containing curatable named-entities.

Most biocurators use rule-based ranking approaches on manually curated data, but these approaches are based on the curators' experience and intuition. That is, the curatable sentences are determined by several named-entities criteria. For example, co-occurrence of targets in the same sentence and the targets were included in the CTD annotation. However, rule-based approaches would suffer the same problem from the non-curatable named-entities. To decrease the effect of non-curatable named-entities, we proposed a curatable sentence classifier that is beneficial in regard to exploring curatable sentences. For example, the first sentence in Fig. 2 is determined to be a non-curatable sentence. There are only two gene named-entities, and the sentence does not contain other chemical or disease named-entities. Being determined as a curatable sentence, the second sentence indicates that 15-deoxy-delta(12,14)-prostaglandin j2 has a real or putative therapeutic role towards lung injury. By classifying the curatable sentences in advance, we can avoid recognizing PARP as a curatable named-entity. When identifying curatable named-entities in text, we not only extract the properties of curatable sentences, such as the semantic features (chemical-disease relations) and their neighbor named-entities (peroxisome proliferator-activated receptor-gamma, rosiglitazone, and bleomycin), but the text features are also used, such as the affixed feature "deoxy," the location feature (the last sentence in the paragraph), and the pattern feature "induce." When biocurators annotate the named-entities in the CTD, the curatable named-entities are usually accompanied with named-entities that have different semantic relationships. For example, ascorbic acid affects the expression of bmp4 mRNA, for which the semantic relationship is found to be a chemical-gene interaction. Thus, this work is fixed on the study of named-entities with different semantic relationships and their variations related to the performance of curatable named-entities recognition. We conducted experiments with combinations of different semantically related features for the purpose of identifying curatable named-entities. The results show that semantically related features are in fact important to the curatable NER task.

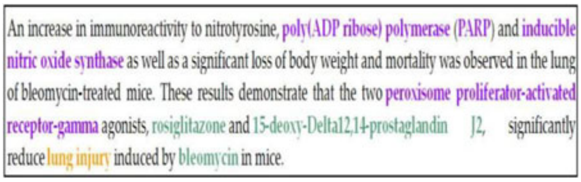
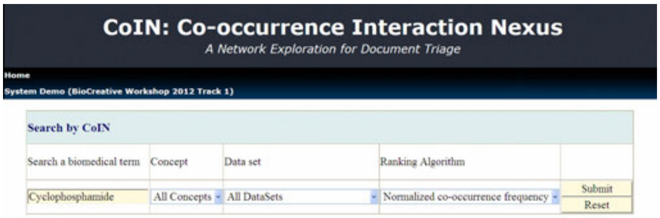

An increase in immunoreactivity to nitrotyrosine, poly(ADP ribose) polymerase (PARP) and inducible nitric oxide synthase as well as a significant loss of body weight and mortality was observed in the lung of bleomycin-treated mice. These results demonstrate that the two peroxisome proliferator-activated receptor-gamma agonists, rosiglitazone and 15-deoxy-Delta12,14-prostaglandin J2, significantly reduce lung injury induced by bleomycin in mice.

Fig. 2. An example of recognizing curatable named-entities.

In a previous work, we proposed a text-mining platform, known as the Co-occurrence Interaction Nexus (CoIN) [17], as shown in Fig. 3. CoIN was developed to distill the entity co-occurrence information from literature and to measure the relationships between entities using a networking approach. Although CoIN has integrated several NER tools, the NER performance still causes problems. We assumed that co-occurrence pairs would be useful in building an automatic curation system. However, we confronted NER problems where the NER rate of disease was low. CoIN


CoIN: Co-occurrence Interaction Nexus

A Network Exploration for Document Triage

Search by CoIN

Search a biomedical term	Concept	Data set	Ranking Algorithm	
Cyclophosphamide	All Concepts	All DataSets	Normalized co-occurrence frequency	Submit Reset

Fig. 3. The input screen of CoIN at http://ikmbio.csie.ncku.edu.tw/coin/home.php.

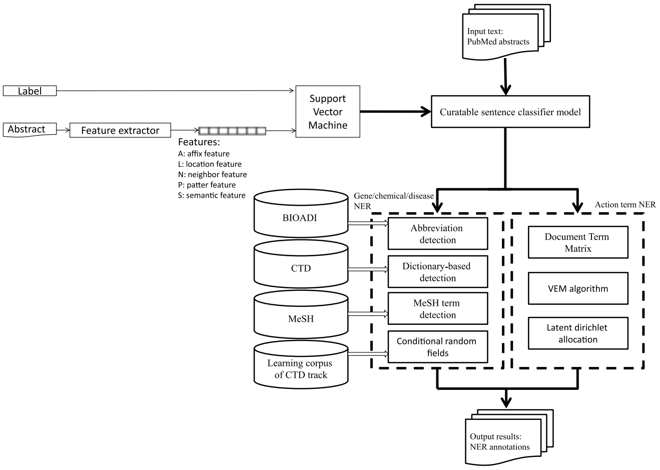


Fig. 4. The workflow of CoINNER.

concerns the relationships between processing entities. In CoIN, PubMed articles are the nodes, while the co-occurrences of gene-disease, gene-chemical, and chemical-disease relationships are the links. By examining the graphical properties, users can gain a global understanding of the likely behavior of the network. Thus, the NER results will have the most influence on the network construction from text. The low recognition rate limits the co-occurrence networks because it generates a great deal of noise. Thus, CoINNER is designed to identify curatable sentences and relieve the problem of the low recognition rate of curatable named-entities. CoINNER not only uses controlled vocabularies, but considers the probability model of annotation corpora. On the other hand, CoINNER collects controlled vocabulary in order to filter the redundancy when the NER task is applied to build co-occurrence networks. CoINNER is capable of convergent services and different NER tasks.

2 Method

For the convenience of biocurators, CoINNER allows users to query genes, diseases, and chemicals. As shown in Fig. 4, PubMed articles are separated into sentences and processed via a curatable sentence classifier. Next, the curatable sentences containing named-entities will proceed to recognize gene names, chemical names, disease names, and action terms. At this stage, CoINNER uses AIIAGMT [18] to identify gene names. We then trained conditional random fields (CRFs) to predict named entities in the articles, and the training patterns are extracted from the CTD. This statistical modeling method is frequently applied in pattern recognition. To tag abbreviations and the MeSH terms, CoINNER employs a dictionary-based method to identify them, and the dictionary is mainly extracted from the Biomedical Abbreviation Definition Identifier (BIOADI) and the MeSH. In the following section, we introduce the curatable sentence classifier, gene/chemical/disease NER and action term NER module, respectively. In the current implementation, CoINNER consists of two NER modules: gene/chemical/disease recognition and action term recognition. The CoINNER has been designed for a RESTful service. Simple XML files specified as BioC are required for input. The output of the system is also generated as the BioC format. A request can choose from one the four categories when the two modules are provided independently. After the two modules accept the BioC XML client request, CoINNER performs NER analysis on the submitted title and abstract, and then returns the BioC XML with annotations representing NER results. In addition, CoINNER normalizes named entities with the identifiers from the MeSH and the CTD.

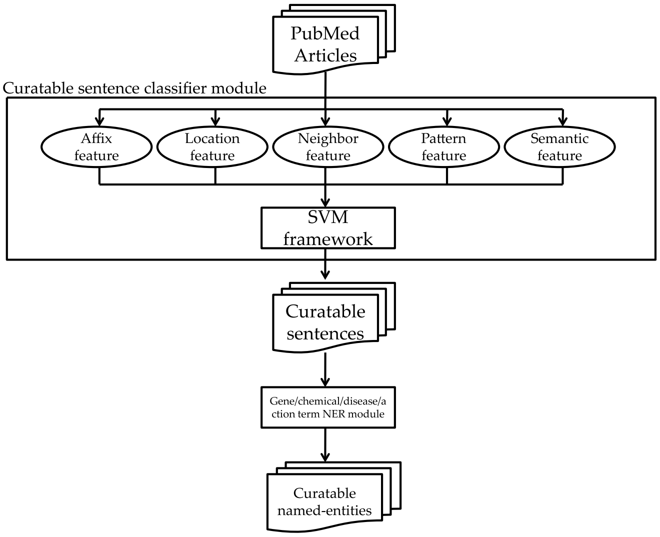


Fig. 5. A workflow for the curatable NER task.

2.1 Curatable Sentence Classifier

The curatable sentence classifier employs a support vector machine (SVM) to distinguish curatable and non-curatable sentences, as shown in Fig. 5. An SVM is an efficient classification algorithm based on statistical learning theory [19]. SVMs have been applied to many classification problems related to supervised learning. Supervised learning usually involves training and testing samples, and training samples include multiple dimension data with many features and a labeled class. Next, an SVM aims to build a model and test the characteristics of the samples. After the SVM framework is established, the model can correctly determine the class labels for unknown instances. The basic idea of SVM is to find a hyperplane that separates two different categories from the sample. In practice, the sample may contain many features resulting in a high-dimensional feature space, so the hyperplane should clearly separate the data into different classes. That is, we can imagine that the hyperplane is a line and that this line forms a border of different categories as large as possible. Another characteristic of an SVM is the ability to solve the classification problem of nonlinear data. When solving the nonlinear data, an SVM projects samples on a higher dimensional space or a feature space. Then, the SVM uses a kernel function to approach the optimal hyperplanes and reduce the empirical classification error. We then introduce the four basic kernel functions [20], [21]: linear, polynomial, radial basis function (RBF), and sigmoid

$$\mathrm{Linear\ kernel}: \mathrm{K}(\mathrm{x_i},\mathrm{x_j}) = \mathrm{x_i^T}\mathrm{x_j}, \tag{1}$$

$$\mathrm{RBF\ kernel}: \mathrm{K}(\mathrm{x_i},\mathrm{x_j}) = \exp(-\gamma\left\|\mathrm{x_i} - \mathrm{x_j}\right\|^2), \gamma > 0, \tag{2}$$

where γ is the width parameter.

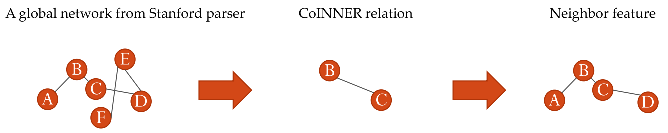


Fig. 6. A toy example of the neighbor feature.

The linear function is the most basic kernel function, and the RBF kernel is more appropriate for practical use. Hence, we used both linear and RBF kernels to build SVM models for curatable sentence classification. In this work, we performed a five-fold cross validation to compare the SVM models with both kernels. In the first module, an SVM curatable sentence classifier is trained with the semantic-related features. After the SVM classification, the curatable sentences proceed to NER modules and retrieve the curatable named entities. PubMed articles are segmented into sentences by the Stanford parser [22] and then proceed to do part-of-speech (POS) tagging. To train an SVM model, we further annotated 100 articles from the training corpus of the BioCreative IV CTD Track. Because the training corpus is provided with interaction pairs, we could mark the sentences that belong to wither curatable or non-curatable relationships. Then, several semantic-related features were applied to represent the characteristics of curatable sentence. For example, "Delirium due to a diltiazem-fentanyl CYP3A4 drug interaction" is classified as a curatable sentence. In this sentence, diltiazem and fentanyl would decrease the activity of CYP3A4, and diltiazem and fentanyl would act as a marker for delirium. There are two different semantic relations (gene-chemical and chemical-disease interactions). By applying the semantic relation of gene-chemical interaction, the sentences including CYP3A4 and other chemicals in the same paragraph can be annotated as curatable. Each feature is illustrated in the following sections.

2.1.1 Affix Feature

Chemical named-entities have their basic affix rules; for example, words staring with "methyl" can be identified as having relative biological functions. That is, we can classify chemical named-entities with some rules such as prefixes and suffixes. We collected organic chemistry glossaries from Wikipedia.

2.1.2 Location Feature

When writing articles, authors used to describe important or key words in the title, topic sentence, and conclusion sentence. We ranked the sentences according to their location of appearance in the paragraph. The articles are divided into N sentences, including title (the first sentence of an article), so the location feature for the title is set as 1/N.

2.1.3 Neighbor Feature

After applying the Stanford parser, we can obtain the nouns and noun phrases from an article. Then, we also can construct named-entity networks, as shown in Fig. 6. For example, A and B are the nouns (noun phrases) that co-occur in the same sentence, and there is an edge between A and B. After the curatable sentence classifier and CoINNER, B and C are annotated as curatable named-entities. Thus, we also extracted A and D. That is, if a sentence contains A, B, C, or D, a neighbor feature is set to 1. The neighbor feature provides feedback that trains SVM models.

TABLE 1
A Verb List Resulting from CRAFT and the Action Term NER

Verb	Frequency
Express	1,233
Observe	809
Compare	668
Describe	621
Perform	599
Develop	597
Form	511
Analyze	364
Differentiate	166
Lose	74

2.1.4 Pattern Feature

The verb in a sentence indicates an action, especially the interaction type between nouns. Here, we briefly collected a verb list from the CRAFT corpus [23] and the action term NER results from the BioCreative IV CTD training corpus, as shown in Table 1. For example, "HCC was induced with diethylnitrosamine and N-nitrosomorpholine." The word "induce" is considered to be a verb that is highly co-related with interaction pairs. On the other hand, we also considered words consisting of experimental terms such as "affymetrix," "probe," and so on.

2.1.5 Semantic Feature

Curatable sentences mean there are at least two different types of named-entities interacting with each other. For example, doxycycline (chemical named-entity) results in decreased expression of MYC (gene named-entity). To describe the semantic feature S included in the sentences, we count the number of entity-entity relationships (EERs) in a single sentence. If EER < 2 in a sentence, $S = 1$; EER $= 2$; $S = 2$; EER $= 3$; $S = 3$; EER $= 4$; $S = 4$; EER > 4, and $S = 5$. It can be observed that most curatable sentences occur in $S = 2$ to $S = 4$.

2.2 Gene/Chemical/Disease NER

This module comprises four components: abbreviation detection, dictionary-based detection, MeSH term detection, and CRFs. For abbreviation detection, we used BIOADI to map the abbreviations in the articles [24]. For dictionary-based detection, we collected the controlled vocabularies of the CTD, which are provided from the BioCreative IV CTD Track website. On detection of MeSH terms, we downloaded the MeSH terms form the National Library of Medicine and matched the named entities with the MeSH terms. CRFs play an important role in CoINNER. Before training CRFs to predict gene, chemical, and disease names, we extracted training patterns from the CTD. We randomly chose 200 articles in the learning corpus of the CTD track,

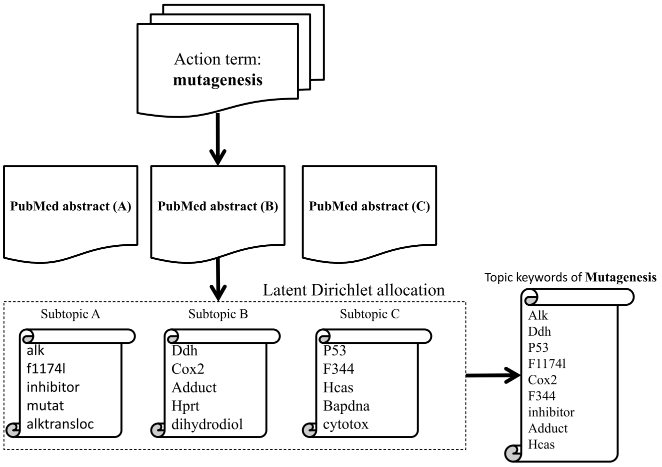


Fig. 7. A toy example of action term NER.

and these articles were further analyzed to determine the training patterns. After retrieving the training patterns, we used CRF++ for the actual implementation (http://crfpp.googlecode.com/svn/trunk/doc/index.html). A probability-based sequence detection CRF model is defined as the conditional probability distribution P(Y | X) of label sequence Y given observation sequence X

$$P(Y|X) = \frac{\exp(F(X,Y))}{\sum_{y'} \exp(F(X,Y'))}, \quad (3)$$

where $y_1, \ldots; y_n$ is a label sequence from Y and x_1, …, and x_n is a token sequence from X.

2.3 Action Term NER

The action term NER is not only involved in information retrieval, but also in higher-level informational reasoning. The process for the action term NER can be considered as where each action term is a main topic with a mixture of various articles (subtopics) in the learning corpus of the CTD track. Therefore, we applied LDA to generate topic keywords for action terms. LDA can be presented as a graphical model for topic discovery [25]. As shown in Fig. 7, the action term "mutagenesis" includes three articles. We then set the number of subtopics as 3 and estimated an LDA model using the variational expectation-maximization (VEM) algorithm. After the LDA model is computed, we can extract the keywords for subtopics. In the CoINNER action term NER, we selected the top 10 keywords for the purpose of representing each subtopic, and we removed the overlapping keywords among the subtopics. Finally, the action terms were annotated by mapping the topic keywords.

2.4 System Interoperability

We developed CoINNER as an advanced tool in the design stage in order to facilitate document triage. However, once users decide to adapt CoINNER into their curation pipeline, it is quite simple for to use exchanging data. For instance, users can use CoINNER to pre-compute the annotation results via a simple request in real-time. When testing CoINNER on a typical modern desktop computer with a Core i5-3550 3.3 GHz CPU and 8 GB RAM, the average required time for processing a PubMed abstract was found to be about 4.2 seconds.

3 Results and Discussion

3.1 Dataset

The training dataset for the triage task for the BioCreative IV CTD Track was distributed to participants (http://www.biocreative.org/tasks/biocreative-iv/track-3-CTD/). The dataset was critical for participants to understand the NER process of the CTD curation; thus, the dataset that consisted of 1,112 articles had been previously triaged and curated by the CTD biocurators [26]. The training dataset was organized into a series of input files that contained all associated curated data in a single BioC XML-based file, which can be found at the following URL (https://gillnet.mdibl.org/~twiegers/bciv/bcIVLearningCorpus.xml), including the PubMed ID, title, abstract, gene, chemical, disease, and action term annotations, and a list of associated curated interactions. The test dataset that consisted of 510 articles was manually curated by the CTD staff. The test dataset contained 1,122 genes, 1,192 chemicals, 943 diseases, 966 chemical/gene-specific action terms, and 3,953 manually curated interactions.

3.2 Evaluation

To assess performance and to make a comparison with our approaches, we use precision, recall, and the F-measure to calculate the performance score. In this work, precision is the fraction obtained when the number of text mined terms is divided by the number of curated term hits. Recall is the fraction obtained when the number of curated terms is divided by the number of curated term hits. The F-measure was calculated as follows:

$$\text{F-measure} = 2 * ((\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision})) \quad (4)$$

A support vector machine is currently the state-of-the-art approach for classification. The curatable sentence classifier is trained by adding features, so we conducted an additional experiment to determine how the curatable sentence classifier would perform in a classification problem. In order to determine the performance of the curatable sentence classifier, we applied five features, such as the affix feature (A), location feature (L), neighbor feature (N), pattern feature (P), and semantic feature (S) to an SVM and the curatable sentence classification process on the BioCreative IV CTD Track training corpus. We used LibSVM [27], which is the most popular SVM implementation package. Then, we added five features to evaluate the contribution of the classifier, as shown in Table 2. To investigate the curatable sentence classifier with the linear and RBF kernel functions, we randomly selected one hundred PubMed abstracts from the BioCreative IV CTD Track, and we manually annotated the sentences with curatable or non-curatable information. Clearly, both kernel functions are equivalent with regard to classifying curatable sentences. However, the SVM-linear model provided greater precision as compared to the SVM-RBF model. Al-though the SVM-RBF model showed better recall, we decided to use the SVM-linear model in order to

TABLE 2
The Performance of the Curatable Sentence Classifier with Different Feature Combinations[1]

	SVM-Linear model			SVM-RBF model		
Config.	Precision	Recall	F-measure	Precision	Recall	F-measure
L+N+S	0.714	0.620	0.664	0.704	0.648	0.675
A+L+N+S	0.775	0.803	0.789	0.765	0.808	0.786
L+N+P+S	0.793	0.754	0.773	0.752	0.783	0.767
A+L+N+P+S	0.826	0.788	0.806	0.782	0.818	0.800

[1] *The following abbreviations are used: A, affix feature; L, location feature; N, neighbor feature; P, pattern feature; and S, semantic feature.*

increase the precision. When investigating the influence of kernel functions, we further analyzed the affix pattern and the pattern feature. By adding these two features, the SVM-linear model performed more precisely than the SVM-RBF model because the SVM-linear model is capable of dealing with spare data and text features, especially title and conclusion sentences. That is, the SVM-linear model is useful when there is prior knowledge about the training data. However, the SVM-RBF model showed better recall than the SVM-linear model. The SVM-RBF model is efficient for providing curatable sentences with neighbor candidates. Using the pattern feature and semantic feature produced better results than not using them because our classifier focuses on the interaction-related properties between genes, diseases, and chemicals.

We examined CoINNER from the two test sets: BioCreative IV_CTD and BC2012CTD. For the BioCreative IV_CTD test set, we excluded the 200 articles for training CRFs and used the remaining articles in the CTD track learning corpus. The BC2012CTD test set, which consisted of 444 articles, was collected from three target chemicals: urethane, phenacetin, and cyclophosphamide. In the training stage, we first evaluated the recall and precision of our modules in finding genes, chemicals, diseases, and action term mentions, as shown in Table 3. Note that the overall is the macro-average of gene/chemical/disease NER. For the training dataset, both of the chemical NER results have a better F-measure, which may have contributed to the fact that approximately half of the target chemicals were found. However, when we consider the tagging behavior of biocurators, CoINNER may provide more opportunities to curate terms intuitively. For the test dataset, we did not analyze and optimize for the submission. According to our training set experiments, we focused on the NER performance recall and the combinations of named-entity pairs. For the submitted run, we did not apply the named-entity normalization because we found that the rate for recognizing named-entities was underestimated. The highest recall and F-measure of the NER module were chemicals, and they achieved 0.885 and 0.619, respectively. It should be noted that the performance of gene/chemical/disease NER for the BC2012 CTD training dataset was superior to that of the BioCreative IV CTD training dataset. As discussed above, the characteristics of the BC2012 CTD training dataset raised the possibility that biocurators might annotate the articles in the specified categories. In our system, the development of the gene/chemical/disease NER is closely tied to the effective use of the co-occurrence approaches.

Table 4 presents the effect of applying CoINNER on the BioCreative IV CTD Track. After building the NER modules with the training dataset, we submitted CoINNER to the BioCreative IV CTD Track 3 organizers' Web service URLs for testing. As shown in Table 4, the overall F-measure of the official result is 0.423, which is inferior to the performance of the training dataset. Therefore, we adjusted CoINNER for named-entity normalization because we did not handle these problems in our original submission for abbreviation and full name mapping. For example, “PMP22” is the abbreviation of “peripheral myelin protein 22.” In our original submission, we submitted both results. However, it would return a false positive with a true positive if PMP22 is an interaction term in this task evaluation. Hence, the adjusted result outperforms the official result at a 0.476 overall F-measure. In this work, we assumed that the curatable sentence classifier can improve the precision of NER results because of the sensitivity of the interaction properties. As a result, CoINNER with the curatable sentence classifier achieved 0.597, 0.725, and 0.523 F-measures for the gene, chemical, and disease NER tasks, respectively. That is, the best average CoINNER F-measure was 0.615. In general, the average F-measures achieved by participants for gene/protein, chemical/drug and disease NER tasks were 0.36, 0.54, and 0.39, respectively [26]. During the test phase of the BioCreative IV CTD Track, ten groups submitted 39 web services

TABLE 3
CoINNER Results Using the Training Dataset

	BioCreative IV_CTD (Training corpus)			BC2012CTD		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Overall	0.456	0.641	0.524	0.469	0.578	0.501
Gene	0.444	0.563	0.496	0.457	0.367	0.407
Chemical	0.450	0.865	0.592	0.477	0.885	0.619
Disease	0.473	0.495	0.483	0.474	0.481	0.477

TABLE 4
CoINNER Results Using the Test Dataset

	BioCreative IV_CTD (Official result)			BioCreative IV_CTD (Adjusted)			CoINNER with Curatable Sentence Classifier		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Overall	0.325	0.677	0.423	0.454	0.525	0.476	0.624	0.619	**0.615**
Gene	0.417	0.581	0.485	0.528	0.418	0.466	0.672	0.537	0.597
Chemical	0.255	0.919	0.399	0.394	0.629	0.484	0.648	0.824	0.725
Disease	0.303	0.532	0.386	0.440	0.527	0.479	0.552	0.496	0.523

that were successfully tested. The average F-measure for the ten groups was 41 percent and ranged from 19 to 54 percent. The results of the CoINNER were significantly superior to those of the participants. Comparing with the best F-measure team for the BioCreative IV CTD Track, CoINNER showed a better F-measure in regard to disease NER because CoINNER focuses on providing curatable interactions in general cases. Note that the best results from the BioCreative IV CTD task were found to be 0.61, 0.74, and 0.51 [9]. However, the best results are not generated by a singular system; we believe that CoINNER with curatable sentence classifier is an integrated system which outperforms other developed systems in regard to the overall NER result.

4 Conclusions and Future Work

In this study, we designed a curatable sentence classifier that retrieves curatable sentences, and then we used dictionary- and CRF-based approaches to develop an NER tool (CoINNER). Our proposed system evaluated the curated terms from the CTD datasets. After applying CoINNER with the curatable sentence classifier, CoIN applies the co-occurrences of sentence structures and the linking activities between biomedical terms, such as genes, chemicals, and diseases, to rank the importance of articles. The system shows how it dealt with the CTD-related NER tasks by the combination of a user defined corpus and the integration of distinct data resources. The system provides users with a valuable ranking result and a customized point of view. Our approaches are helpful for biocurators to enhance the efficiency of biocuration in an automated system. CoIN begins with the automatic identification of named-entity recognition (CoINNER) and links in each named-entity. After constructing heterogeneous co-occurrence networks from the combinations of different named-entity pairs, we computed the co-occurrence frequency and network centralities of co-occurrence networks. If an article has more named-entity pairs, the article has a higher priority to be curated. We tested CoINNER with the test data in the CTD track of BioCreative IV, and CoINNER demonstrated its ability to annotate named-entities. The experiments with the test data showed that CoINNER achieved a 0.476 overall F-measure with an average response time of 4.2 seconds. Combined with the curatable sentence classifier, CoINNER obtained a 0.615 overall F-measure, which exhibits a balanced system by which to handle NER tasks.

In the future, we hope to utilize the sentence structure of named-entity pairs in CoINNER. Although we have investigated gene-disease, gene-chemical and chemical-disease named-entity pairs, we believe that more detailed research focusing on the integration of multiple NER tools will improve the performance of CoINNER.

References

[1] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee, "Big data: The future of biocuration," *Nature*, vol. 455, pp. 47–50, Sep. 4, 2008.

[2] A. Zouaq and R. Nkambou, "A survey of domain ontology engineering: Methods and tools," in *Advances in Intelligent Tutoring Systems*. vol. 308, R. Nkambou, J. Bourdeau, and R. Mizoguchi, Eds. Berlin, Germany: Springer, 2010, pp. 103–119.

[3] O. Bodenreider and R. Stevens, "Bio-ontologies: Current trends and future directions," *Briefings Bioinformatics*, vol. 7, pp. 256–274, Sep. 1, 2006.

[4] L. Yao, A. Divoli, I. Mayzus, J. A. Evans, and A. Rzhetsky, "Benchmarking ontologies: Bigger or better?," *PLoS Comput. Biol.*, vol. 7, p. e1001055, 2011.

[5] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nat. Biotechnol.*, vol. 24, pp. 537–44, May 2006.

[6] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *Amer. J. Hum. Genet.*, vol. 78, pp. 1011–25, Jun. 2006.

[7] A. Neveol, R. Islamaj Dogan, and Z. Lu, "Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction," *J. Biomed. Inform.*, vol. 44, pp. 310–8, Apr. 2011.

[8] T. C. Wiegers, A. P. Davis, and C. J. Mattingly, "Collaborative biocuration–text-mining development task for document prioritization for curation," *Database (Oxford)*, vol. 2012, p. bas037, 2012.

[9] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, "BioCreative-IV virtual issue," *Database*, vol. 2014, p. bau039, Jan. 2014.

[10] A. P. Davis, T. C. Wiegers, R. J. Johnson, J. M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, C. G. Murphy, and C. J. Mattingly, "Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database," *PLoS One*, vol. 8, p. e58201, 2013.

[11] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, pp. 3191–3192, Jul. 15, 2005.

[12] P. Corbett and A. Copestake, "Cascaded classifiers for confidence-based chemical named entity recognition," *BMC Bioinformatics*, vol. 9, no. Suppl. 11, p. S4, 2008.

[13] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program," in *Proc. AMIA Symp.*, 2001, pp. 17–21.

[14] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *J. Biomed. Inform.*, vol. 40, pp. 288–299, Jun. 2007.

[15] C. H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H. Y. Kao, and Z. Lu, "Accelerating literature curation with text-mining tools: A case study of using PubTator to curate genes in PubMed abstracts," *Database (Oxford)*, vol. 2012, p. bas041, 2012.

[16] S. Kim and W. J. Wilbur, "Classifying protein-protein interaction articles using word and syntactic features," *BMC Bioinformatics*, vol. 12, no. Suppl. 8, p. S9, 2011.

[17] Y. Y. Hsu and H. Y. Kao, "CoIN: A network analysis for document triage," *Database (Oxford)*, vol. 2013, p. bat076, 2013.

[18] C. N. Hsu, Y. M. Chang, C. J. Kuo, Y. S. Lin, H. S. Huang, and I. F. Chung, "Integrating high dimensional bi-directional parsing models for gene mention tagging," *Bioinformatics*, vol. 24, pp. i286–94, Jul. 1, 2008.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[20] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Nat. Taiwan Univ., Taipei, Taiwan, http://www.csie.ntu.edu.tw/cjlin/ papers/guide/guide.pdf, 2003.

[21] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, pp. 113–126, Jan. 2004.

[22] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 455–465.

[23] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. Baumgartner, K. Cohen, K. Verspoor, J. Blake, and L. Hunter, "Concept annotation in the CRAFT corpus," *BMC Bioinformatics*, vol. 13, p. 161, 2012.

[24] C. J. Kuo, M. H. Ling, K. T. Lin, and C. N. Hsu, "BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature," *BMC Bioinformatics*, vol. 10, no. Suppl. 15, p. S7, 2009.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[26] T. C. Wiegers, A. P. Davis, and C. J. Mattingly, "Web services-based text mining demonstrates broad impacts for interoperability and process simplification," presented at the *4th BioCreative Challenge Evaluation Workshop*, Bethesda, MD, USA, 2013.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 1–27, 2011.

Yi-Yu Hsu received the MS degree from the Department of Information and Learning Technology at the National University of Tainan, Taiwan, R.O.C., in 2008. He is currently working toward the PhD degree in the Department of Computer Science and Information Engineering at the National Cheng-Kung University, Tainan, Taiwan, R.O.C.

Hung-Yu Kao received the BS and MS degrees in computer science from the National Tsing Hua University, Hsinchu, Taiwan, in 1994 and 1996, respectively. In July 2003, he received the PhD degree from the Electrical Engineering Department, National Taiwan University, Taipei, Taiwan. He was a postdoctoral fellow of the Institute of Information Science (IIS), Academia Sinica, from 2003 to 2004. He is currently an associate professor of computer science and information engineering at the National Cheng Kung University. His research interests include web information retrieval/extraction, search engine, knowledge management, data mining, social network analysis, and bioinformatics. He has published more than 60 research papers in refereed international journals and conference proceedings. He is a member of the IEEE and ACM.