Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus

Zhong Huang and Xiaohua Hu

***Abstract*—Named Entity Recognition (NER) has been an active research fields in biomedical text mining. In the past years, much attention has been focused on semantic types related to protein, gene, and other named entities in biology domain. Human disease named entity recognition in literatures, however, has not received much attention. Comparing the NER solutions targeting protein/gene named entities, existing machine learning solutions lacks same level of precision and recall for disease named entity recognition. The development of machine learning based NER for disease named entity is largely focused on local features of tokens in the sentence, by integrating its linguistic, orthographic, morphological, local contextual characteristics. In this paper, we utilized the sentence level semantic contextual information as one of discriminative features for disease NE recognition. Our method takes advantage of semantic types related to disease in UMLS metathesaurus by fuzzy dictionary lookup. The results show promises to improve the performance of current disease NER methods.**

***Index Terms*—Biomedical concept, disease, named entity, named entity recognition, NE, NER, semantic type, machine learning, conditional random fields, CRF.**

I. Introduction

Named Entity Recognition (NER) refers to the computational method to automatically recognize named entities (NE) in natural language documents, e.g. to relate it to a named entity (NE) in the domain of interest. For biomedical domain, a NE is defined as a single word term or multi-words phrase that denotes a biomedical object, for instance a protein, gene, disease, or drug with which a semantic hierarchy is associated.

NER in biomedical text mining is particularly challenging. It is evidenced by the fact that many alias, different naming conventions, abbreviations, variety of organisms may refer a same protein/gene with different terms, or a term may refer to different biologically different entities. For example, named entity p53 may refer to a protein name in one context, but may also be used to denote the molecular weight of a protein with 53 Kd in another context. To tackle those problems different approaches have been applied on NER using rule based, dictionary matching based, and machine learning based techniques. With rapid accumulation of biomedical literatures published in thousands of journals, many new terms and spelling variations of existing terms have emerged. For those terms the rule based and dictionary based approaches lacks prediction power. Machine learning based approaches, on the other hand, have been demonstrated as the most robust method for biomedical NER due to its capability of handling high dimensional discriminative vector features in text processing and prediction of new terms or variations based on learned patterns. To train a high performance and reliable NER model, it is important to fully capture features surrounding the word in the context. For the past years, biomedical NER systems have been developed using linguistic characteristics of the word (word stemming and lemmatization), the orthographic features (formation of the word such as presence of upper case, symbols, digits etc), the morphological features (suffixes/prefixes, char n-grams, and word shape), and local context features (word window and conjunctions) [1]. Some systems also integrated exact dictionary matching to recognize named entities in a domain specific dictionary. The binary encoding of the feature set is used as input for the machine learning algorithm to train the NER model, along with the human annotation of NE mentions in the training dataset [2]. In recent years, the Conditional Random Fields (CRFs) has been used as supervised machine learning method for several high-performance NER systems due to its relaxation on feature independence assumptions hence the advantage of handling high dimensional arbitrary feature sets over other machine learning methods such as Hidden Markov Models (HMMs), Maximum Entropy Markov Models (HEMMs), and Support Vector Machines (SVMs).

Despite commonly used feature sets as described above have been utilized successfully by many public NER systems, those approaches didn't take into account the global semantic information, such as correlated concepts on the sentence level. It is intuitive to think such global semantic information can be used as discriminative features to further disambiguate word therefore improve the NER performance.

For the past years, much attention has been focused on NER of gene and protein products, while little work has been conducted on disease NER. In this paper, we present a new method to extract concept features on the sentence level for CRF based NER machine learning. The paper is organized as follows. First we briefly introduce the related works, Conditional Random Fields and selecting of feature set for NER machine learning. We then introduce a new method to construct the semantic concept feature using semantic types from UMLS thesaurus. Finally the experiment evaluation for our approach is given in results and discussion section.

A. Related Works

Several text mining systems have been implemented for

Manuscript received September 16, 2013; revised November 5, 2013.

The authors are with School of Information Science and Technology, Drexel University, Philadelphia, USA (e-mail: zhong.huang@drexel.edu).

biomedical NER tasks using different approaches. Those approaches, in summary, can be categorized into following four categories.

1) Dictionary-based approach is the most straightforward approach that tries to find all NE from text by looking up the dictionary. Some nomenclatures have been extensively applied on biomedical text mining. The HUGO Nomenclature for instance, provides more than 21,000 human gene entries [3]. The Swiss-Prot, the UniProt database containing more than 180,000 protein records has also been frequently used. The Bio Thesaurus collects comprehensive compilation of several million human protein and gene names mapped to UniProt knowledgebase entries using cross-reference in iProclass database [4]. Unlike machine learning based approach, one major advantage of dictionary based approach is that it has external database identifier (ID) built-in for each entry, thus provides external metadata annotation to the extracted names. However, it suffers from several limitations including false positive caused by name ambiguity, false negative cause by spelling variations and synonyms, and inability to cover newly created names. In addition, it heavily depends on creation and curation of lexicon for the specific domain which may consist of millions of entries and is very labor intensive. To address aforementioned spelling variation issue, Tsuruoka et.al used approximate string searching and variant generator methods to achieve a significant improvement of F-measure (10.8%) on GENIA copora evaluation as compared with exact matching algorithms [5].
2) Rule-based approach can better deal with word orthographic and morphological structures, as compared with dictionary based approach. In [6] a method using surface clue on character strings was presented to identify core terms followed by handcrafted patterns and rules to concatenate adjacent words as named entity. The rule based approach largely depends on the domain specific named entities with common orthographic or morphologic characteristics. Thus makes it difficult to extend to other domains since the handcrafted rules are often domain specific and cannot be applied to a new domain due to different naming conventions.
3) Machine learning approaches are most frequently used and have achieved the best performance in BioCreative II gene/protein NER tasks. Different supervised machine learning methods including HMMs [7], [8], SVM [9], MEMMs [10], CRF [11], and Case-based reasoning [12] have been used in NER systems. In addition to supervised methods that utilize only the annotated text corpora, in order to solve data sparseness issue which often encountered when using large feature set on an relatively small training dataset, some semi-supervised methods are also presented recently to take advantage of large size of un-annotated text corpora. Such semi-supervised machine learning include semi-CRFs [13], semi-SVMs [Bennett 1999], ASO [14], and FCG [15]. One critical step of machine learning approaches is to select the most discriminative feature set that represent the NE. Commonly used features include orthographical word formation patterns, morphological patterns, part-of-speech POS tagging, lemmatization, token window, and conjunction of contextual features.

B. Data Set

Biotext corpus was originally annotated for disease and treatment mentions [16] and is part of Biotext Project at UC Berkley. The corpus was obtained from MEDLINE 2001 and contains 3655 annotated sentences.

To extract the concepts from sentences we used semantic types of UMLS metathesaurus. It defines a comprehensive hierarchical tree of semantic network to represent all concepts in the UMLS metathesaurus as well as their relationships. This semantic network currently contains 133 semantic types and 54 relationships. Fig. 1 shows the UMLS semantic network hierarchy related to disease.

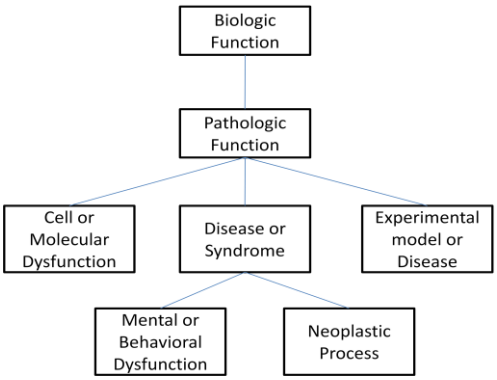

Fig. 1. UMLS semantic network disease related semantic type hierarchy.

In addition to extracting concept by fuzzy matching from sentence using UMLS metathesaurus, we also included exact dictionary matching to add the semantic feature using a manually curated human disease dictionary containing 25,944 entries.

C. Feature Extraction

Fig. 2 shows the system architecture for disease NER. The corpus was first pre-processed by tokenization and lemmatization before feature extraction. Following [17], we used feature set consisting of POS, lemma, orthographical and morphological features (patterns for word capitalization, letter and digit combinations, prefixes and suffixes). Numbers were normalized by converting digits to single digit "0".

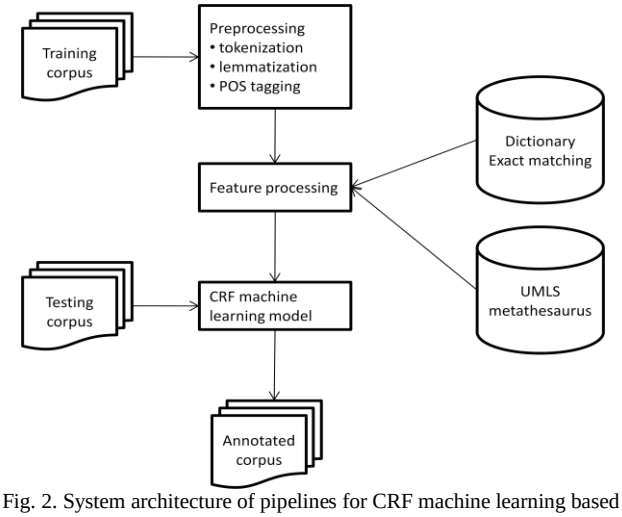

Fig. 2. System architecture of pipelines for CRF machine learning based disease NER.

Exact dictionary matching using a disease dictionary was utilized to add biomedical knowledge semantic information to the feature. If a word in the sentence is matched with the dictionary entry, it is labeled as a feature for machine learning.

One limitation of exact dictionary matching for NER is that it often gives false negative for spelling variations and newly created terms in the text. Moreover, it is heavily dependent on the availability of domain specific dictionary which is not easily portable to other domains. For this reason, we propose a new method that utilizes semantic types of UMLS metathesaurus to extract disease related concept from text and use it as one of discriminative features, along with features described above, for NER machine learning.

We used the approximate dictionary lookup algorithm that was first presented in [18] to capture the significant word in the text, in contrast to capturing all words of the concept, and map it to the ontology term, e.g. UMLS semantic concept.

Let concept $c = \{s_1, s_2, s_3, \ldots, s_n\}$, where s_1-s_n are variant concept names that belong to c. Let $N(w)$ denotes number of concepts whose variant names contain word w.

The relative significance score of word w to the concept c is defined as:

$$I(w,c) = max\{I(w,s_j) \mid j \le n\} \quad (1)$$

where: $I(w,s_j) = \begin{cases} 0 & w \notin s_j \\ \frac{1/N(w)}{\sum_i 1/N(w_{ji})} & w \in s_j \end{cases}$

A huge significant scores matrix containing normalized words as rows and concepts as columns were built using UMLS Metathesaurus [18] and stored as sparse matrix for efficient retrieval. In equation 1 shown above, the w_{ji} denotes the word at *i-th* row which is found in concept s_j at *j-th* column.

The concept lookup algorithm used rule-based pattern matching to search the word boundary and extract the concept term from text. In this study we used the default threshold score of 0.95 and the maximum number of skipped words of 1 which have been shown to give the best results for UMLS based biological concept extraction.

The word that is mapped to an UMLS concept is then filtered by its semantic type shown in Fig. 1. Only those concepts with semantic type of "DISEASE OR SYNDROME" are kept. The word with filtered semantic type is assigned a label and encoded as a new binary feature for model training at next step. The algorithm for this conceptual semantic feature generation is shown in Fig. 3.

D. Conditional Random Fields

In this study, we used conditional random fields (CRF) machine learning algorithm which has been proved to be a high performance method for label sequence problem. In [11] CRF was proposed as an undirected graphical model and the conditional probability of output nodes can be calculated based on other designated input nodes. The model use conditional probability for inference by defining a single log-linear distribution over label sequences of Y, given the observation sequence of X. It combines the idea of Hidden Markov Model (HMM) which deals with sequences problem, and Max-Entropy (ME) that utilizes many correlated features. However, CRF maximize the conditional probability $p(y \mid x)$ directly, while HMM maximize the joint probability $p(x, y)$. In the meantime, it avoided label bias problem compared to Maximum Entropy Markov Models (MEMMs), and is capable of handling arbitrary features with relaxed independence assumption as compared to HMMs. In text mining fields, the sequence of words is regarded as special case of linear chain of output nodes.

```
Algorithm concept feature engineering

Input:
1. sentence consisting of n words
2. semantic types of concept for extracted concept filtering

Output:
UMLS concept semantic type tagged sentence

SET sentence S = {w1, w2, ... , wn} where w1-wn is the pre-processed word
SET D = {d1, d2, ..., dm} where d1-dm is the semantic type of the concept to be kept
Initiate array of semantic type for w1-wn (ArrayW) and SET each value to 0
Initiate the list FL for final filtered concept names
Find next starting word ts
k = 0
C = {c | t ∈ T(c) } // T(c) is the set of words in concept c
For each c ∈ C, Sc = I(ts, c) // Sc is the significant score for word ts to concept c
WHILE next word t is not boundary word AND k < skip
    N = {c | t ∈ T(c) ∧ c ∈ C}
    IF N = Φ Then k = k + 1
    Else
        C = N
        For each c ∈ C
                Sc = Sc + I(t, c)
        End If
WhileEnd
C = {c | Sc > threshold ∧ c ∈ C }
If |C| > 0 then
    Return concept name and candidate concepts c ∈ C
End If

For each c ∈ C
    Get its UMLS term id tui
    Get the semantic type for tui
    If semantic type of the tui ∈ D
        Add the concept c to final list FL
For each word w in c ∈ FL
    Get its position index p (0≤p ≤n)
    Set ArrayW[p] = 1
Return array of semantic type for w1-wn (ArrayW)
```

Fig. 3. Algorithm for extracting concept from sentence and generate the binary concept feature for machine learning.

Let define the undirected graph $G = (V, E)$ such that a node $v \in V$ and the random variable represents an element Yv of Y which is indexed by the vertices of G. The (Y, X) is a conditional random field when conditioned on X, and the random field Yv obeys the Markov property with respect to G. e.g. $p(Yv \mid X, Yw, w \neq v) = p(Yv \mid X, Yw, w \sim v)$ where $w \sim v$ denotes the neighbors in G. Therefore the CRF is a random field globally conditioned on the observation X.

For text labeling problem, let $s = \{o_1, o_2, \ldots, o_T\}$ be the observed sequence of words from a sentence with length s. Let S be a set of states in a finite state machine with each associated a label. The conditional probability of a state sequence $s = \{s_1, s_2, \ldots, s_T\}$ is calculated as:

$$P_\wedge(s \mid o) = \frac{1}{Z_o} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (2)$$

where $f_k(s_{t-1}, s_t, o, t)$ is a feature function with λ_k as weight that can be learned during model training. The Zo is a normalization factor of all state sequences which is used to

sum up all conditional probabilities to 1 and is calculated as:

$$Z_o = \sum_s \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k (s_{t-1}, s_t, o, t) \right) \quad (3)$$

The objective function to be maximized in CRF model training is the log-likelihood of the state sequences given observation sequences:

$$L_\wedge = \sum_{i=1}^{N} \log\left(P_\wedge \left(s^{(i)} \mid o^{(i)} \right) \right) - \sum_k \frac{\lambda_k^2}{2\sigma^2} \quad (4)$$

where ($s^{(i)} \mid o^{(i)}$) is the empirical distribution of training data. The L-BFGS algorithm is used for CRF parameter estimation and can be treated as a black-box optimization procedure.

In a nutshell, given a sentence of *n* words for named entity labeling problem, we want to predict the tag *T* for a given word *W* using linear-chain CRF such that $P(T \mid W) = \frac{1}{Z} \exp(\theta \cdot F(T))$ and maximize the weight $\theta \cdot F(T)$.

We used the 2-order CRF implemented in Mallet toolkit for our experiment [19].

II. Results and Discussion

A. Experimental Settings

For the experiment of human disease NER task the golden-standard Biotext corpus was used which contains 3655 annotated sentences [13]. The sentences without close xml tag were removed which result in a final corpus of 3580 annotated sentences. Due to relatively small dataset, the 5 × 2 fold cross-validation was used for evaluation. The test is executed for 5 iterations of 2-fold cross-validation. Compared with 10 fold cross-validation, it is more powerful in terms of detecting real system performance differences rather than the biased splitting of testing data.

Precision (*P*), recall (*R*), and *F*-measure (*F*-score) were used as evaluation metric shown in formula below:

$$\begin{aligned} P &= TP / (TP + FP) \\ R &= TP/ (TP + FN) \\ F\text{-score} &= (2 \times P \times R) / (P + R) \end{aligned} \quad (5)$$

where *TP*, *FP*, and *FN* are numbers of true positive, false positive, and false negative.

B. Tokenization and Part-of-Speech (POS) Tagging

We used a simple tokenization method to tokenize the sentence. For POS tagging, we experimented two different POS taggers implemented in Dragon Toolkit [20], namely Hepple tagger and Med Post tagger. Med Post tagger is a POS tagger [21] specifically designed for biomedical text as compared with the more generic Hepple tagger. As shown in Table I, our results show an improvement in *F*-score by 1.23 using Med Post tagger over Hepple tagger when the disease specific dictionary is used. When a larger dictionary combining both non disease specific dictionary and disease specific dictionary was used, it slightly decreases precision, recall, and *F*-score of Med Post tagger.

TABLE I: Evaluation with Hepple Tagger and MedPost Tagger

	P (%)	R (%)	F-score (%)
Hepple Tagger + non disease specific dictionary	61.90	47.79	54.28
Hepple Tagger + disease dictionary	63.29	48.21	54.72
MedPost Tagger + disease dictionary	64.93	49.15	55.95
MedPost Tagger + combined dictionary	64.45	48.80	55.54

Non disease specific dictionary contains biological entities not specific to disease. Disease dictionary contains 25,944 entries of manually curated human disease names. The combined dictionary contains both non disease dictionary entries and the disease dictionary entries.

C. Named Entity Encoding Scheme

As discussed above in CRF section, NER can be modeled as a sequence labeling problem. Let $x = \{x_1, x_2, \ldots, x_n\}$ be the sequence of tokens for the input sentence, the problem is to determine the output sequence of labels $t = \{t_1, t_2, \ldots, t_n\}$ such that $t_i \in L$ (set of labels) for $1 \leq i \leq n$. The output label consists of two parts, e.g. the named entity type and its positional information.

We first compared 3 named entity position encoding scheme, namely IO, BIO, and BIOEW. Our results shown in Table II suggest the more complex coding schemes do not necessarily increase the F-score for Biotext corpus NER task. The IO encoding scheme gives the slightly better *F*-score than BIO and BIOEW schemes. This is in agreement with the finding in [17] that uses the BioCreative II corpus for gene/protein NER task. In this paper, the IO setting is retained for our experiments.

TABLE II: Results of Evaluating Different Entity Encoding Scheme on Biotext NER Task

	P (%)	R (%)	F-score (%)
IO	61.90	47.79	54.28
BIO	63.40	47.13	54.07
BIOEW	63.11	46.61	53.61

Hepple tagger and non disease specific dictionary were used.

D. Effect of Semantic Concept Feature

As shown in Table II our preliminary experiment using exact disease dictionary matching indicates the biomedical knowledge can improve the performance of disease NER. We further experimented the effect of using concept semantic type as a new feature for disease NER. Table III shows results using the disease concept semantic type, e.g. "DISEASE OR SYNDROME" (type-1). The result without concept semantic type feature (type-0) is used as baseline for comparison.

TABLE III: RESULTS OF EVALUATING EFFECT OF CONCEPT SEMANTIC TYPES AS FEATURE FOR DISEASE NER

	P (%)	R (%)	F-score (%)
Type-0	64.93	49.15	55.95
Type-1	65.98	49.67	56.67

Type-1 is "DISEASE OR SYNDROME" semantic type. Type-0 denotes no concept semantic feature added.

Table III shows that by adding "DISEASE OR SYNDROME" semantic type as feature to train the CRF model achieves overall 0.72 increase of F-score, with 1.05 and 0.52 increase in precision and recall.

Three NER systems for disease recognition using the Biotext corpus and 5 × 2 cross-validation was reported in [17]. Comparing with their results, our semantic type feature based method gives the highest F-score of 56.67 (BANNER: 54.84, ABNER: 53.44, and LingPipe: 51.15). This is due largely to the increase of recall (BANNER: 45.55, ABNER: 44.86, LingPipe: 47.50). The performance of disease NER using Biotext by different systems are poor, as compared with performance on gene and protein NER using BioCreative II gene mention task. This could be due to several reasons. First, the Biotext golden-standard corpus is considerably small (3655 sentences versus 20,000 sentences for BioCreative II corpus), which is more likely to cause the data sparseness and out-of-vocabulary (OOV) issue. Secondly, unlike Biotext that has only one annotation, the BioCreative II gene mention task provides an alternative annotation.

III. CONCLUSIONS

This paper presents a new method of utilizing biomedical knowledge by both exact matching of disease dictionary and adding semantic concept feature through UMLS semantic type filtering to improve the human disease named entity recognition by machine learning. By engineering the concept semantic type into feature set, we demonstrated the importance of domain knowledge on machine learning based disease NER. The background knowledge enriches the representation of named entity and helps to disambiguate terms in the context thereby improves the overall NER performance.

REFERENCES

[1] J. K. Jong and C. Park, "Named entity recognition," in *Text Mining for Biology and Biomedicine*, J. McNaught, Ed. 2006.

[2] J. L. O. D. Campos and S. Matos, "Biomedical named entity recognition: A survey of machine-learning tools," in *Theory and Applications for Advanced Text Mining*, S. Sakurai, Ed. InTech, 2012.

[3] R. G. Cotton, V. McKusick, and C. R. Scriver, "The HUGO mutation database initiative," *Science*, vol. 279, no. 5347, pp. 10–1, Jan. 1998.

[4] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: A web-based thesaurus of protein and gene names," *Bioinformatics*, vol. 22, no. 1, pp. 103–5, Jan. 2006.

[5] Y. Tsuruoka and J. Tsujii, "Improving the performance of dictionary-based approaches in protein name recognition," *J. Biomed. Inform.*, vol. 37, no. 6, pp. 461–70, Dec. 2004.

[6] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: Identifying protein names from biological papers," *Pac. Symp. Biocomput.*, pp. 707–18, Jan. 1998.

[7] N. Collier, C. Nobata, and J. Tsujii, "Extracting the names of genes and gene products with a hidden Markov model," in *Proc. the 18th Conference on Computational Linguistics*, 2000, vol. 1, pp. 201–207.

[8] G. D. Zhou, "Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid.," *Int. J. Med. Inform.*, vol. 75, no. 6, pp. 456–67, Jun. 2006.

[9] S. Jonnalagadda, T. Cohen, S. Wu, H. Liu, G. Gonzalez, J. Siddhartha, C. Trevor, W. Stephen, H. F. Liu, and G. Graciela, "Using empirically constructed lexical resources for named entity recognition," *Biomed. Inform. Insights*, vol. 6, no. Suppl. 1, pp. 17–27, Jan. 2013.

[10] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy Markov models for information extraction and segmentation," pp. 591–598, Jun. 2000.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Machine Learning - International Workshop*, 2001, pp. 282–289.

[12] M. L. Neves, J.-M. Carazo, and A. Pascual-Montano, "Moara: A Java library for extracting and normalizing gene and protein mentions," *BMC Bioinformatics*, vol. 11, pp. 157, Jan. 2010.

[13] G. Mann and A. McCallum, "Efficient computation of entropy gradient for semi-supervised conditional random fields," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.*, 2007, pp. 109–112.

[14] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Dec. 2005.

[15] Y. Li, X. Hu, H. Lin, and Z. Yang, "A framework for semisupervised feature generation and its applications in biomedical literature mining," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 2, pp. 294–307.

[16] B. Rosario and M. Hearst, "Classifying semantic relations in bioscience texts," in *Proc. 42nd Annu. Meet. Assoc. Comput. Linguist*, 2004.

[17] R. Leaman, G. Gonzalez, L. Robert, and G. Graciela, "BANNER: An executable survey of advances in biomedical named entity recognition," *Pac. Symp. Biocomput.*, pp. 652–663, Jan. 2008.

[18] X. Zhou, X. Zhang, and X. Hu, "MaxMatcher: Biological concept extraction using approximate dictionary lookup," *PRICAI 2006 Trends Artif. Intell.*, pp. 1145–1149, 2006.

[19] A. K. McCallum. (2002). MALLET: A machine learning for language toolkit. [Online]. Available: http://mallet.cs.umass.edu

[20] X. Zhou, X. Hu, and X. Zhang, "Dragon toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining," in *Proc. 19th IEEE Int. Conf. Tools with Artif. Intell. 2007*, 2007, vol. 2, pp. 197–201.

[21] L. Smith, T. Rindflesch, and W. J. Wilbur, "MedPost: A part-of-speech tagger for bioMedical text," *Bioinformatics*, vol. 20, pp. 2320–2321, 2004.

Zhong Huang received Master Degree in information science and Physiology from Drexel University and Beijing University Medical Center. He is currently a Ph.D candidate at Drexel University, Philadelphia. His research interests are in the areas of data mining, text mining, and bioinformatics.

Xiaohua Hu is a professor and the director of Data Mining and Bioinformatics Lab at the College of Information Science and Technology, Drexel University. He obtained his Ph.D Degree from University of Regina in 1995. His current research interests are in data/text/web mining, big data, bioinformatics, social network analysis, healthcare informatics, rough set theory and application.