Zhang HT, Huang ML, Zhu XY. A unified active learning framework for biomedical relation extraction. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 27(6): 1302–1313 Nov. 2012. DOI 10.1007/s11390-012-1306-0

A Unified Active Learning Framework for Biomedical Relation Extraction

Hong-Tao Zhang (张宏涛), Min-Lie Huang (黄民烈), and Xiao-Yan Zhu (朱小燕), *Member CCF*

State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

E-mail: mathzhanghongtao@163.com; {aihuang, zxy-dcs}@tsinghua.edu.cn

Received October 17, 2011; revised May 25, 2012.

Abstract Supervised machine learning methods have been employed with great success in the task of biomedical relation extraction. However, existing methods are not practical enough, since manual construction of large training data is very expensive. Therefore, active learning is urgently needed for designing practical relation extraction methods with little human effort. In this paper, we describe a unified active learning framework. Particularly, our framework systematically addresses some practical issues during active learning process, including a strategy for selecting informative data, a data diversity selection algorithm, an active feature acquisition method, and an informative feature selection algorithm, in order to meet the challenges due to the immense amount of complex and diverse biomedical text. The framework is evaluated on protein-protein interaction (PPI) extraction and is shown to achieve promising results with a significant reduction in editorial effort and labeling time.

Keywords biomedical relation extraction, active learning, unified framework

1 Introduction

One of the most motivations for the biomedical text mining is the exponential growth of the biomedical literature and the urgent need of biologists to seek information more accurately and efficiently[1]. In addition, the growth in new publications is still in great expansion[2], which makes it difficult for biologists to keep up with the new information hidden in those new publications. Therefore, automatic text mining methods are required to facilitate the biomedical information acquisition.

Since biomedical relations, such as protein-protein interaction (PPI), gene-gene interactions, play an important role in understanding biological processes[3], biomedical relation extraction is a very important research topic in the field of biomedical text mining. The major goal of relation extraction is to discover the relations embedded within sentences, paragraphs, or entire documents. Over the past years, significant progress has been made in biomedical relation extraction by adopting supervised machine learning methods[4-17], which essentially represent each candidate relation pair and its context by a feature vector or the weights of features which are learned from labelled training data. However, such methods tend to suffer from the bottleneck, since labeling large training data is very expensive and even unrealistic. Therefore, how to achieve promising results with a small amount of labelled data still remains a challenge.

Recently, active learning comes as a framework to reduce the labeling effort in supervised learning, and has shown good performance when the labelled data is in short supply[18]. The key idea of active learning is to iteratively select a small set of unlabelled examples to be labelled and added into the training data, in order to maximally improve the classifier's performance. Specifically, a typical active learning system is composed of two parts, that is, a learning module and an example selection module which work iteratively. In each iteration, the learning module trains a model based on the current training data，while the example selection module selects the most informative unlabelled samples for manual labeling to enrich training data. Unfortunately, to the best of our knowledge, there is still limited work on biomedical relation extraction by active learning. Moreover, due to the intrinsic characteristics of biomedical text, such as the complexity and diversity of language used in biomedical domain and the different annotation schema employed in different

Regular Paper

The work is supported by the National Natural Science Foundation of China under Grant No. 60973104 and the National Basic Research 973 Program of China under Grant No. 2012CB316301.

groups[5], current methods that only focus on the example selection strategy cannot cope with more complex tasks. Considering a learning model used in PPI extraction task which only accesses to a small amount of labelled training data, because other labelled data require complex or expensive manual construction, it is quite possible that the model is to extract PPIs using incomplete feature description because of the restriction of extracting features from limited labelled data. In other words, enriching and selecting most informative features from selected examples during the active learning procedures is also a very important step. Therefore, the systematic design of active learning for biomedical relation extraction is an essential task.

In this paper, we propose a unified active learning framework for biomedical relation extraction. In particular, we systematically study several issues involved in active learning process, including a strategy for selecting informative data, a diversity data selection algorithm, an active feature acquisition method, and an informative feature selection algorithm. In informative data selection stage, two simple but effective selection strategies, *maximum uncertainty based strategy* and *density-based strategy*, are employed to select informative examples in groups. In diversity data selection stage, an efficient heuristic selection algorithm is employed to enforce selected data to have no duplicates, with the purpose of reducing computational requirements and making it feasible for large-scale application with thousands of examples. In active feature acquisition stage, rule-based methods are proposed to enrich features from the selected data, in order to allow the learner to express complete feature information for the entire population. In informative feature selection stage, a feature selection method is proposed to identify the most relevant features between training and test data. The experimental results on protein-protein interaction (PPI) extraction show that the proposed framework is practical and effective. More importantly, our proposed framework is generic and may be applicable to the extraction of all biomedical relations.

The remainder of this paper is structured as follows. In Section 2, we discuss previous approaches, including supervised machine learning methods for biomedical relation extraction and recent active learning application in some fields. In Section 3, a brief description of biomedical relation extraction is given. In Section 4, the unified active learning framework is presented. Section 5 shows the experimental results supporting the efficiency of our framework; finally, the conclusion and future work is given in Section 6.

2 Related Work

In biomedical field, researchers are usually interested in PPIs, gene-gene interactions and protein-disease interactions. The major goal of relation extraction is to discover the relations embedded within sentences, paragraphs, or entire documents[19]. Currently, the most popular relation extraction methods are based on supervised machine learning. These methods can be broadly characterized into feature-based or kernel-based, depending on the manner in which samples are represented. In feature-based methods, candidate relation pairs are represented by a feature vector, which usually includes bag-of-word features, part-of-speech (POS) tagger features and parser-related features. Katrenko and Adriaans made use of dependency parsing information and employed Bayes net, Naïve Bayes and K-nearest neighbor to detect PPI[4]. Miwa *et al.* designed a rich feature vector with three types of features, including bag-of-words features, shortest dependency path features and graph features made from dependency tree, in order to express important information for PPI extraction[5]. Yang *et al.* developed a system for PPI extraction based on support vector machines (SVMs) and the link grammar parser. The set of features in that paper includes surface word, keyword, protein name distance and link path features[6]. Li *et al.* used Feature Coupling Generalization (FCG), a recently proposed semi-supervised learning strategy, to learn an enriched feature representation of local contexts in sentences from millions of unlabelled samples[7]. Landeghem *et al.* reported extensive study of feature selection for bio-molecular event extraction, where they not only analyzed the contribution of different feature types, but also investigated the most important features within one specific type. They reported that the features expressing syntactic information about the trigger words (called trigger features in that paper), lexical information about triggers and the bag-of-words features appear to be highly relevant and include practically no irrelevant features, while the part-of-speech taggers of the words on the syntactic trees as well as the trigrams appear to be much less informative in general[8]. Bui *et al.* proposed a hybrid approach to extract protein-protein interactions. They firstly applied semantic rules to partition the dataset into subsets according to its semantic properties and extract candidate PPI pairs from these subsets; secondly, they introduced enhanced feature sets for use with an SVM classifier to classify these extracted PPI pairs[9]. In addition, van Landeghem *et al.*[10], Fayruzov *et al.*[11], Miyao *et al.*[12], and Niu *et al.*[13] also studied individual impact of a variety of feature types on the PPI extraction task. In kernel-based methods, candidate relationship pairs are encoded as structural representations such as bag-of-words, word-sequence, parse trees or dependency graphs to measure the similarity between them. Erkan *et al.* defined one

kernel based on cosine similarity and another on edit distance between the paths between the protein names to extract protein interaction sentences[14]. Kim *et al.* designed four kernels: predicate, walk, dependency, and hybrid kernels to encapsulate the information of sentential structures for relation prediction[15]. Airola *et al.* developed all-path graph kernel to make use of full, general dependency graphs for representing the sentence structure[16]. Segura-Bedmar *et al.* used a shallow linguistic kernel for drug-drug interaction extraction[17]. In summary, both the feature-based and the kernel-based methods achieved state-of-the-art performance on benchmark datasets. Most of the above methods are based on supervised machine learning, which means large enough amounts of labelled data are required to train the learning model. However, the manual construction of training data is proven time and resource consuming, it would be nice to achieve promising result with a small amount of labelled data.

Active learning is well motivated in many modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain[18]. In comparison with passive learning that trains models with pre-collected large training data, active learning is able to select the most representative data in an iterative manner based on the model learned in each iteration. From a practical point of view, active learning concerns about the selection strategy for informative examples, the batch-mode setting and its variants, and the strategy for feature enrichment when examples have incomplete feature descriptions. Active learning has been widely explored in many kinds of research field for its capability of reducing human annotation effort, including multimedia research community[20], learning to rank on web search[21], image retrieval[22], machine translation[23]. In BioNLP (natural language processing of biology text) domain, active learning has been employed in sequenced-based PPI predication[24]. Unfortunately, this work only focuses on the example selection strategy, rather than the general active learning framework analysis. In this paper, we address this problem by systematically studying active learning in the biomedical domain, where we not only focus on example selection, but also discuss the acquisition of new features and diverse example selection, in order to design an effective and practical active learning method.

3 Biomedical Relation Extraction Task

As mentioned in Section 2, the goal of biomedical relation extraction is to detect occurrences of relations between a pair of entities of given types. While the type of the entities is usually very specific (e.g., genes, proteins or drugs), the type of relations may be very general (e.g., any biochemical association) or very specific (e.g., a regulatory relation).

Take PPI extraction as an example. Let us consider the following sentence containing three protein names (shown in italic): *NAT*1 binds *eIF*4*A* but not *eIF*4*E* and inhibits both cap-dependent and cap-independent translation. This sentence contains three protein pairs, namely *NAT*1-*eIF*4*A*, *NAT*1-*eIF*4*E*, and *eIF*4*A*-*eIF*4*E*. Generally, a protein pair is a positive example if the original sentence expresses an interaction between members of this pair, and a negative example if they just co-occur in the sentence. Therefore, there is only one positive example, namely, *NAT*1-*eIF*4*A*, while the other two examples are negatively hidden in the above sentence. Based on these, the task of protein interaction extraction is setup as a binary classification task: each feature vector corresponds to a pair of proteins and it is classified as positive pair or negative pair.

Formally, the relation extraction extractor is a function to a set of triples, $\{\langle Ent1, rel, Ent2\rangle\}$, where $Ent1$ and $Ent2$ are biomedical entities and rel is a textual fragment indicating the relation between the two entities. The extractor should produce one triple for every relation stated explicitly in the text, but is not required to infer implicit facts. At present, we usually assume that all relational examples are stated within a single sentence.

4 Active Learning Framework

Active learning attempts to overcome the labeling bottleneck by iteratively selecting a small amount of unlabelled data to be labelled by an "*oracle*" (e.g., a human annotator), aiming to achieve high performance using as few labelled examples as possible, and minimize the cost of obtaining labelled data. To design a practical and effective active learning framework, there are four practical issues to be considered[18]:

1) How evaluate the informativeness of unlabelled examples?

2) How improve the diversity of the selected data when employ the batch-mode setting during active learning process?

3) Should new features be generated from the selected data to overcome the incomplete feature descriptions?

4) How identify the most relevant features based on the entire population?

In this section, a unified active learning framework is presented to address these problems mentioned above. Our proposed framework mainly consists of four components: an information data selection strategy based on *maximum entropy* and *density*, a diversity data selection algorithm incorporating a diversity measure, rule-

based active feature acquisition methods, and an informative feature selection algorithm.

4.1 Informative Data Selection

All active learning methods involve selecting informative unlabelled examples. The key point is how to measure the informativeness of an unlabelled example, and select a new example with maximal informativeness to augment the training data. In this paper, we employ two simplest and most commonly used strategies, the density-based strategy and the maximum uncertainty based strategy.

Maximum Uncertainty Based Strategy. The strategy implies that the current classifier has the least confidence in its classification of this example. The well-known *entropy* is a good uncertainty measurement widely used in active learning[18]:

$$x^* = \arg\max_x \Big\{ -\sum_i p_M(y_i|x) \log p_M(y_i|x) \Big\}, \quad (1)$$

where x^* means the most informative instance according to the entropy measurement, $p_M(y_i|x)$ stands for the posterior probability under model M, x is the input example (in this paper x is the candidate relation pair), y_i ranges over all possible class labels. For binary classification, $y_i \in \{0, 1\}$. Therefore, the method is equivalent to select the examples with a class posterior closest to 0.5, so does it in this paper.

As shown in the above equation, this strategy iteratively selects a single new example from a set of unlabelled examples, queries the corresponding class label and then performs retraining of the current model. However, sometimes the time required to induce a model is slow or expensive, especially in biomedical domain when using advanced natural language processing (NLP) tools. To reduce computational time for training, it might be necessary to select batches of new training examples instead of single example, namely, the batch-mode setting, as follows:

$$D(x^*) = \{x^* | f_{\min} \leqslant x^* \leqslant f_{\max}\}, \quad (2)$$

where $f_{\min}$ and $f_{\max}$ are two prediction threshold values. In this paper, we empirically define these two threshold values with the purpose of selecting more and more informative examples. For instance, $f_{\min}$ and $f_{\max}$ are usually assigned as 0.45 and 0.65. When there are not enough examples in this interval (less than 150 in this paper), we will appropriately expand this interval, in order to select enough uncertainty examples.

Density-Based Strategy. The strategy is another batch-mode setting strategy. The main idea of the density-based strategy is that informative instances should be those which are "representative" of the underlying distribution (i.e., inhabit dense regions of the input space)[18]. In this paper, we implement this strategy by using the K-means algorithm. In each iteration, we fix the number of examples to be labelled. The selected examples are distributed across the clusters in proportion to the size of the cluster. In each iteration, we select the examples that are closest to the cluster's centroid.

4.2 Diversity Data Selection

In general, the batch-mode active learning is more efficient when a parallel labeling instance is available, e.g., a number of labels can be determined at the same time by an experimental test procedure. However, from the practical point of view, we have to consider the turnaround time between obtaining a newly labelled example from the human annotator to generating the next example. It is essential to make full use of turnaround time, rather than waste the editorial resources by presenting copies of the selected examples for labeling. Therefore, we need to improve the diversity of the selected examples, in order to select the most informative examples while without duplicate ones. In this paper, we employ a heuristic algorithm based on cosine distance[25]. The algorithm can be modified to be a two-step process. First, we construct an initial selected examples set by using the maximal entropy with batch-mode setting (using (1) and (2)). Then, we adopt a heuristic selection algorithm incorporating a diversity measure, in order to do data diversity selection. The details are shown in Algorithm 1.

Algorithm 1. Diversity Data Selection

Input: n, the threshold value of the size of the resulting diversity dataset.

Output: The resulting diversity set H.

1: Initialize $H = \varnothing$.

2: Construct the initial selected examples set $D(x^*)$ using maximal entropy with batch-mode setting.

3: **For** any two examples x_i, x_j in $D(x^*)$, calculate the cosine distance:

$$\text{cosine}(x_i, x_j) = \frac{\boldsymbol{w}(x_i) \cdot \boldsymbol{w}(x_j)}{\|\boldsymbol{w}(x_i)\| \|\boldsymbol{w}(x_j)\|},$$

where, $\boldsymbol{w}(x_i)$ and $\boldsymbol{w}(x_j)$ are the feature vectors of x_i and x_j, respectively.

4: Add x_i, x_j into H if these two examples have the maximal cosine distance.

5: Reset $D(x^*) = D(x^*) \backslash \{x_i, x_j\}$.

6: **While** $n^* < n$ (n^* is the size of H) **Do**

$x_k = \arg\max_{x_k \in D(x^*)} \{\arg\min_{x_l \in H} \{\text{cosine}(x_k, x_l)\}\}$,

$H = H \cup x_k$,

$D(x^*) = D(x^*) \backslash \{x_k\}$

End.

7: Output H.

Note that, when the number of selected examples is relative small (less than 150 in this paper), we suggest that it should be better to make full use of the selected examples, instead of doing diversity data selection.

4.3 Active Feature Acquisition

As we continually select new examples into the training data, it is quite possible that new features that are hitherto unseen are also being made available to the classifier. For example, consider a classifier for PPI extraction and suppose that the training data do not contain the keyword "complex" and we restrict ourselves to the original keyword feature space. It should be noted that keyword feature is a kind of significant feature in PPI extraction. In this paper, we refer the keywords list generated in [26] as the original keyword feature space. Now even if we select the examples that contain the keyword "complex" for labeling, we still risk missing a very important feature for this classifier by the keyword feature space restriction. In our active learning framework, we attempt to minimize the risk of losing out on important features by generating features from the selected examples, namely, the active feature acquisition. At present, keyword features, shortest dependency path features, and lexical pattern features significantly contribute to biomedical relation extraction[12,26]. At the same time, we notice that previous study has reported that the generalization of POS tagger patterns from lexical patterns is crucial for a text mining framework, as it enables extraction and prediction of events concerning previously unpublished entities[27]. Therefore, in this paper, we construct the POS tagger pattern features from the original lexical pattern. Then, we try to enrich these kinds of features from the selected true examples through active feature acquisition. To be brief, we give simple description for each kind of features, shown in Table 1. Then, three rule-based methods are designed for the acquisition process for these three types of features.

4.3.1 Active Acquisition of POS Tagger Pattern Features

We present the rule-based method as follows:

1) Generate POS taggers sequences for each true example, including the i ($i = 4$) taggers (if there is any) before the first entity, i taggers (if there is any) after the second entity, and all taggers between the two entities. The entities themselves are represented by a special token "*Ent*".

2) Filter the "illegal" sequences. If a sequence has neither verb tag nor noun tag, reject it; if the last tag of a pattern is *IN* or *TO*, reject it; if the left neighborhood of a *CC* tag is not equal to the right one in the pattern, reject it.

3) Remove useless tags from each sequence. The useless tags include *JJ*, *JJS* (superlative adjective), *JJR* (comparative adjective), *RB*, *RBS* (superlative adverb), *RBR* (comparative adverb) and *DT*, which are given detailed description in [26].

4) Format these sequences as the soft matching patterns. For example, the sequence "*Ent NN IN Ent*" is formatted as "*Ent* $*$ *NN* $*$ *IN* $*$ *Ent*".

5) Re-rank these patterns according to the frequency that each pattern appears in test data. Select top n patterns as the candidate pattern features.

4.3.2 Active Acquisition of Keyword Features

The method for keyword features is given as follows:

1) For each true example, extract nouns and verbs at the corresponding position, including *before* (four tokens (if there is any) before the first entity), *between*, and *end* (four tokens (if there is any) after the second entity).

2) Filter the obvious "noise" and remove those tokens that are already in the original keywords list by the *oracle*.

3) Re-rank these keywords according to the frequency that each keyword appears in test data. Select top n keywords as the candidate keyword features.

Table 1. Feature Description

Feature	Examples	Description
Keyword	*active; regulate activation; interaction*	The original keyword features in this paper include noun keywords and verb keywords, which are all reproduced from [26].
Shortest dependency path	*pobj* $*$ *prep* $*$ *nn* *nn* $*$ *pobj* $*$ *prep* $*$ *nn*	The shortest dependency paths are all extracted from true relation pairs. The tokens, such as "*pobj*" and "*prep*", are dependency relations. In this paper, these paths are used as the syntactic patterns, in which the asterisk in these paths indicates that any word can be skipped, namely, the soft matching.
POS tagger pattern	*Ent* $*$ *VB* $*$ *Ent* *NN IN* $*$ *Ent* $*$ *IN* $*$ *Ent*	All the original POS tagger patterns are reproduced from [3]. In this paper, all the lexical words in lexical patterns are replaced with their POS tags. The biomedical entities, such as gene, protein, are all replaced by a special token "*Ent*". Similarly, these POS tagger patterns employ the soft matching.

Finally, the new shortest dependency path features are directly extracted from these true examples.

By using the active feature acquisition, we can exploit new features and compensate for feature distributions' difference between the training data and the test data, which is better than restricting ourselves to the initial feature space.

4.4 Informative Feature Selection

Through the active feature acquisition, the number of candidate features grows significantly. In theory, more features should provide more discriminating power[28]. However, in fact, due to the limited amount of training data, it is common knowledge that a large number of features are either irrelevant or redundant with respect to the class concept.

Take PPI extraction task as an example. The dataset analyzed here is the BioInfer corpus①, where 10% randomly chosen data for training and the other 90% as the test data. The original feature space includes keyword features, shortest dependency path features, and POS tagger pattern features. As shown in Fig.1, although the three types of features are all significant in PPI extraction task, most of features are not available in both training and test data. Therefore, it is very important to do informative feature selection, with the purpose of identifying the most relevant features and exhibiting the maximal predictive performance.

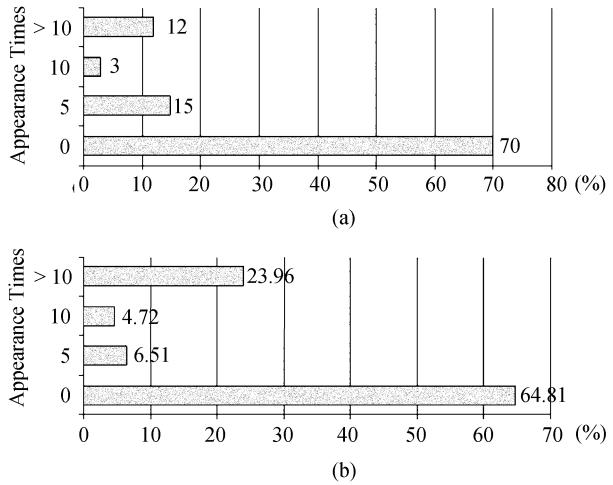

Fig.1. Coverage of features. The legends at the left side of the figure means that one feature appears zero time, or no more than five times, or more than five times and no more than ten times, or more than ten times, respectively. (a) Features coverage in training data. (b) Features coverage in test data.

The key aspect of feature selection is to measure the *relevance* of features. Here, we define two measures in order to identify the most relevant features between training and test data. To find the common relevant keywords between training and test data, we calculate the relevant score based on the following metric:

$$s(w_i) = \Big(\frac{p_{\text{training}}(w_i) + p_{\text{test}}(w_i)}{2}\Big) \times e^{(-|p_{\text{training}}(w_i) - p_{\text{test}}(w_i)|)}, \tag{3}$$

where, $p_{\text{training}}(w_i)$ and $p_{\text{test}}(w_i)$ are the probabilities of the candidate keyword w_i occurring in the training and test data, respectively.

If w_i has high score, which indicates that w_i occurs frequently and similarly in both training and test data, then it should be considered as a common relevant keyword feature.

To identify the common relevant patterns, including syntactic patterns and POS tagger patterns, for each candidate pattern, we calculate its relevant score based on the metric modified from AutoSlog-TS[29]:

$$s(p_i) = \begin{cases} e^{Acc(p_i)} \times \log_2(Freq(p_i)), & \text{if } Freq(p_i) > 1, \\ e^{Acc(p_i)} \times 0.1, & \text{if } Freq(p_i) = 1, \\ e^{Acc(p_i)} \times 0.01, & \text{if } Freq(p_i) = 0, \end{cases} \tag{4}$$

where, $Freq(p_i)$ is the frequency of the pattern p_i observed in test data, $Acc(p_i)$ is the accuracy of the pattern p_i in training data. For example, if a pattern p_i totally matches n_i true PPI pairs, and the number of total true PPI pairs in the training data is n, then $Acc(p_i) = n_i/n$. As mentioned above, metric $s(p_i)$ aims to identify the patterns that are precise in training data and satisfied frequently in test data.

Based on the above two metrics, we summarize the feature selection algorithm as Algorithm 2:

Algorithm 2. Informative Feature Selection

Input: F_I, the initial feature set; F_G, the generated feature set; λ, the threshold value of relevance score.

Output: The resulting feature set F.

1: Initialize $F = \varnothing$.
2: **For** each feature F_i in $F_I \cup F_G$, calculate the relevance score s_i, according to (3) and (4), respectively.
3: **If** $s_i \geqslant \lambda$, $F = F \cup \{F_i\}$.
4: Output F.

It should be noted that F_I is firstly extracted from the initial training data. As the active learner iteratively selects examples from the entire population into the training data, F_I is also iteratively updated, in order to minimize the risk of losing out on important features. At the same time, our feature selection methods

①http://mars.cs.utu.fi/PPICorpora/eval-standard.html

are induced from the entire population (rather than just from selected examples). This is because models induced from all available data have been shown to be superior to models induced when examples with missing values are ignored[30].

4.5 Framework Overview

Our framework is shown in Fig.2. Instead of learning from a large pool of labelled data, the framework starts by few labelled examples. More importantly, the classifiers iteratively augment their training data with a limited number of examples from the entire population, and choose the most relevant features generated from the selected examples into the feature space, in order to furthermore improve the learning model.

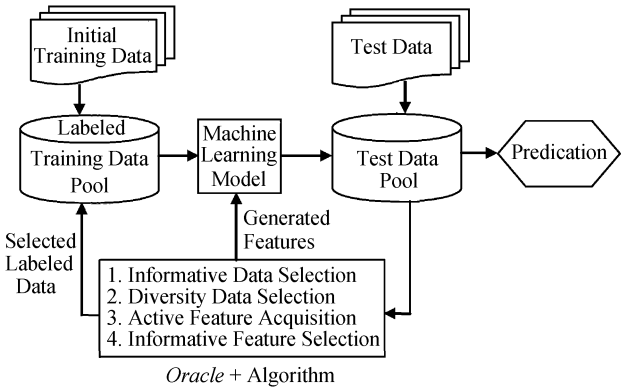

Fig.2. Unified active learning framework.

The basic classifier is constructed by following the method reported in [31], which is based on SVMs with linear kernels. Except for the three types of features described in Table 1, another four types of features are furthermore incorporated in the classifier. The first one is negation keyword features, which are used to exclude those sentences that contain no relation pairs due to the use of negation words. The second one is window POS tagger features, which extract the POS taggers from the words at the corresponding position, including *before*, *between* and *after*. The third one is shortest path POS tagger features. For a given protein pair, we first obtain the shortest path between the two proteins and then extract the POS tags of the nodes in the path as a feature. For example, we first obtain a shortest dependency path $\{Ent * nsubj * Ent\}$ for the lexical path $\{Protein1\ bind\ Protein2\}$; then the shortest POS tagger feature for this path is $\{nn * veb * nn\}$. The fourth one is the dependency relation features, which use the corresponding dependency relation set for each shortest path as a feature. For example, we first obtain the shortest dependency relation path for a given protein pair, such as the path {*prep-pobj-conj-pobj-conj-amod*}. Then we break up the path and obtain the corresponding dependency relation set for this path {*prep*, *pobj*, *conj*, *amod*}. Each dependency relation in the set corresponds to one dimension in the feature space.

Specifically, the original keyword features are reproduced from the mined keyword in [26], the initial shortest dependency path features are extracted from all the true examples in initial training data, and the POS tagger patterns are reproduced from lexical patterns reported in [3]. Instead of the compact feature representation reported in that method, we apply more flexible feature representation in this paper. Considering that the shortest dependency path features adopt soft matching, a shortest path could match more than one syntactic pattern, thus the dependency path feature for one example contains less than five matched paths from the true path set, so does the POS tagger pattern features.

In addition, a divide-and-conquer approach is designed in that method: if the target relation pairs contain less than four tokens, only keyword features, negation keyword features, POS tagger pattern features, and window POS tagger features are used; otherwise, all the features are adopted as described before.

5 Experimental Results and Discussion

As a general extraction framework for biomedical relations, our method can be applied to a wide range of relation extraction applications. Since PPI is central to all the biological processes and structural scaffolds in living organisms, in this paper, we focus on PPI extraction from text to demonstrate the properties and effectiveness of our method.

5.1 Data and Evaluation Measures

All data used in this paper is shown in Table 2. The AIMed corpus consists of 225 abstracts, 200 of which contain annotated human gene and protein interactions. Another 25 abstracts contain protein names but do not describe any interactions. The BioInfer corpus contains 1 100 sentences describing protein-protein interactions. HPRD50 contains sentences that were extracted from a subset of 50 abstracts, referenced by the Human Protein Reference Database (HPRD) and annotated with protein names and interactions between them. The LLL corpus consists of 76 sentences describing interactions concerning Bacillus subtilis transcription. Both HPRD50 and LLL contain relatively short hand-picked sentences with a simple syntactic structure. In this paper, we use these data with a unified formation. More detailed description can be found in [32].

Table 2. PPI Data

Corpus	Positive Pairs	All Pairs
AIMed	1 000	5 834
BioInfer	2 534	9 653
HPRD50	163	433
LLL	335	330

To classify those data, we use the SVM implementation from LIBSVM[②], where the underlying yet effective linear kernel is selected for this binary classifier. Also, the Weka Package[③] is used to implement the K-means clustering. The final predictions are evaluated by the common golden standard evaluation measures, including precision, recall and F-score (harmonic of precision and recall). In this paper, we only report the macro biggest F-score to compare different methods.

5.2 Single Corpus Evaluation

In this subsection, we choose the biggest corpus BioInfer for single corpus evaluation, where 10% randomly chosen data are used as initial training data and the other 90% as the pool data for each round of the active learning. Furthermore, all of the test data are randomly divided into five sub-corpora, namely, we do five iterations during our active learning procedure. We also conduct a baseline that trains a model on the 10% initial training data and directly test the 90% data by this model. Both the number of initial training data and the original feature space are identical. The results are summarized in Table 3.

Table 3. Results on Single Corpus

Method	Precision (%)	Recall (%)	F-Score (%)
Baseline	61.74	34.39	44.17
Density	46.51	57.50	51.42
Uncertainty	47.69	57.00	51.93

It should be noted that *density* refers to using a density-based strategy to select the informative examples, while *uncertainty* refers to using an uncertainty-based strategy. Both methods do active feature acquisition, informative feature selection, and diversity data selection (if needed). As illustrated in the above table, the lower baseline results further prove that different data distributions and low coverage of features in training or test data (shown in Fig.1) significantly degrade the performance, especially when the training dataset is small. After using active learning, both methods achieve considerable improvements on recall and F-score. These results show the effectiveness of the active learning strategies in obtaining better performance when adding a relatively small amount of labelled data (the detailed discussion about the number of added labelled data points will be presented in the following subsection). We think that higher recall means better extraction of correct examples, which establishes a good foundation for the further development of extraction techniques. It is also observed that the active learning methods obtain lower precision than the baseline. This is mainly because the classifier has not learned well enough for these selected examples, especially for the negative examples which may confuse the classifier.

Then, the effectiveness of different paradigms is investigated during the active learning process. We conduct the leave-one-out experiments, including only using IDS (informative data selection), using IDS and AFA (active feature acquisition), using IDS + AFA + IFS (informative feature selection), and IDS + AFA + IFS + DDS (diversity data selection). Table 4 summarizes the experimental results. To be brief, we only study the uncertainty-based strategy in IDS stage.

Table 4. Results with Different Paradigms

Paradigm	Precision (%)	Recall (%)	F-Score (%)
Baseline	61.74	34.39	44.17
IDS	44.36	53.87	48.65
IDS + AFA	42.88	60.65	50.23
IDS + AFA + IFS	48.01	56.30	51.82
IDS + AFA + IFS + DDS	47.69	57.00	51.93

As shown in Table 4, we observe that: 1) AFA outperforms the method only using IDS, since AFA incrementally generates features from these selected examples, which allows the classifier to request more complete feature information to improve the predictive model. 2) The method using IFS furthermore achieves better performance with fewer features than methods with all features, since IFS selects the most informative features to obtain during training, rather than randomly or exhaustively acquiring all new features for all training examples, which is capable of eliminating noisy and irrelevant dimensions. 3) Although the method using DDS outperforms the method not using DDS, the improvement appears to be quite small. This result suggests that selecting diverse examples leads to more effective learning. The performance using DDS is pretty close to the method without using DDS, even obtaining a slight improvement on F-score and recall, suggesting that DDS can select the diverse examples during the learning process. In other words, we can achieve promising results with less labelled data, which is very

②http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

③http://www.cs.waikato.ac.nz/ml/weka/

important for practical applications. In the following subsection, we will give a detailed discussion.

5.3 Practical Considerations

So far, most active learning researches have focused on mechanisms for informative data selection from the learner's perspective. In essence, the recent active learning work attempts to address the problem that can machines learn with fewer training instances, and shows good performances subject to some assumptions. For example, it often assumes that the cost for labeling selected examples is either free or generally expensive, which is not true in many real-world situations. In addition, in biomedical text mining field, annotation is known to be very expensive and time-consuming, due to the complex and diverse biomedical text. Therefore, we think that a practical and effective active learning framework should satisfy the two issues: robust and economical.

The first one is that the designed active learning framework should be feasible for large-scale application with several thousands of examples, which means the performance of this method should always maintain a stable growth trend, rather than dramatically different or even less efficient. We show the learning curves of different paradigms on different test data in Fig.3, in which we report the maximal F-score improvement appears in each fold for each paradigm (compared to the baseline in Table 3). Similarly, we only discuss the uncertainty-based strategy in IDS stage.

As shown in the figure, because of the differences of

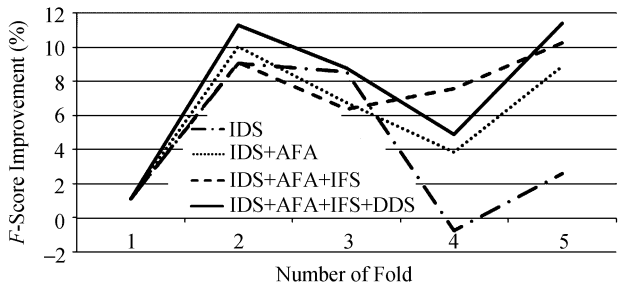

Fig.3. Learning curves of different paradigms on different test data.

data in each fold, the performances of each paradigm varies with different folds, while all the four paradigms are able to achieve noticeable improvement in F-score, except for the IDS in the fourth fold. Furthermore, we observe that the improvement obtained by the IDS paradigm in the fourth fold has a negative value, which means the performance is slightly smaller than that of the baseline. In fact, the improvement made by IDS paradigm varies dramatically with different folds. This is likely because when we do not do AFA, the existing feature space does not express enough information for these new selected examples. That the classifier does not learn the characteristics of new examples well enough and therefore has more confusion as we continually select new uncertainty examples. When we employ another three paradigms, the curves always show a stable growth trend compared with the baseline, because AFA and IFS can explore more and more highly relevant features for these new selected data, which makes the feature space express much better information for the unknown relation extraction. At the same time, DDS can furthermore select the most representative examples, decreasing the possibility of adding noise data (i.e., the so-called less relevant negative examples). Therefore, our unified active learning framework with different paradigms is robust.

The second one is that the active learning framework should take the computational time and the turnaround time into account, in order to avoid wasting the editorial resources by presenting duplicates of the selected examples to be labelled. Table 5 shows the number of examples to be labelled in different folds when using the methods with DDS and without DDS, respectively. Note that the method using DDS selects the equal number of positive and negative examples into the training data in each fold, in order to avoid the class imbalance distribution between positive and negative. As shown in Table 5, we observe that although the number of examples to be labelled decreases sharply, the performance using DDS is even slightly better than that the one without DDS, which means that the classifiers remain valid after using DDS paradigm to those selected examples. In other words, DDS can furthermore select most representative examples from the initial selected

Table 5. Number of Examples to Be Labelled in Different Folds

Paradigm	Class	Fold Number				Total	Performance
		1	2	3	4	Number	F-Score(%)
Without DDS	Positive+Negative	259	333	343	312	1 247 + 835	51.82
	Positive	95	153	126	92		
With DDS	Positive+Negative	190	168	142	104	604 + 835	51.93
	Positive	95	84	71	52		
Common method	Positive + Negative			–		9 653	54.79

examples, so that reduce the number of examples to be labelled.

Furthermore, we compare the active learning results with results using common method. The common method employs the same SVM with a linear kernel used as the baseline in Table 3, but it uses a different proportion on training and test examples. It randomly chooses 80% data from BioInfer for training and 20% data for test. Then, we perform a 5-fold cross validation on this corpus. As shown in Table 5, although the active learning results are evaluated on 90% BioInfer corpus, and are relative lower than the common results, we achieve considerable reduction on editorial effort and labeling time, proving that our method is more economical to deal with the large scale of biomedical data. It should be noted that the "Total Number" row includes the 10% initial labelled training data.

Finally, we present the learning curves of different selection strategies when using the percentage increase of the labelled data during the learning process. First, we randomly choose 10% of data (containing 235 positive examples and 610 negative examples) from the BioInfer corpus as initial training data, the other 45% of the data (containing almost 1 000 positive examples and 3 100 negative examples) as the pool of labelled data for each round of the active learning, and the rest of data (containing almost 1 300 positive examples and 3 500 negative examples) as the test data. In each round, we choose a fixed percentage increase of the examples from the pool of the labelled data, and we add these data into the initial training data. Then, we train a model on these data, and predict the test data by this model. Similarly, we also select equal number of positive and negative examples into the training data in each fold, in order to avoid the class imbalance distribution between positive and negative. In this experiment, we use three selection strategies, including a density-based strategy, an uncertainty-based strategy, and a random-based strategy. Among of them, the former two strategies also employ AFA, IFS, and DDS during the learning process, while the latter one does not use any paradigms. The initial feature spaces are identical for these three methods. Fig.4 shows these learning curves (in terms of F-score).

As illustrated in Fig.4, when we add only 400 examples, the density- and uncertainty-based methods achieve considerable improvements on F-score, while the random-based method only obtains a slight improvement. When we add 1 600 examples (almost 40% of the total labelled data), the density- and uncertainty-based method reach their best performances, which are pretty close to the upper limit 54.18% (the upper limit is obtained by adding 100% of the labelled data). During the whole learning process, the random-based method shows a relative slow growth trend. The learning curves furthermore prove that our unified active learning framework is practical.

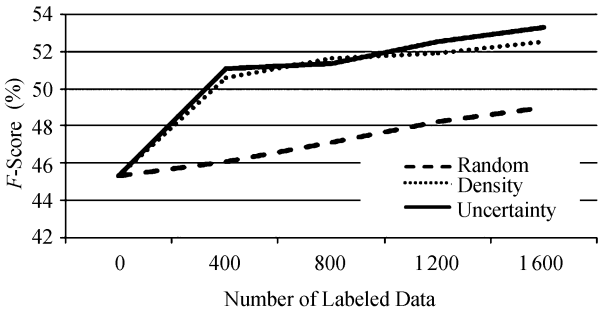

Fig.4. Learning curves of different percentage increases of labelled data.

5.4 Comparison Results

In biomedical domain, it is known that the performances cannot be compared directly because of the differences in corpora and the parsers used in data preprocessing[12]. To compare with other methods, we conduct cross-corpora evaluation, where we use two small corpora HPRD50 and LLL as the initial training data, and two relative large corpora AIMed and BioInfer as the pool data for each fold of the active learning. Similarly, both AIMed and BioInfer are randomly divided into five sub-corpora. In other words, we do five iterations during our active learning procedure. We choose the method reported in [16] for comparison, since we use the same parser and same corpora. All the results are summarized in Table 6. The results in Table 6 demonstrate that our methods outperform others on AIMed and BioInfer corpora.

Table 6. Comparison Results

	AIMed		BioInfer	
	Airola[16]	Our Results	Airola[16]	Our Results
HPRD50	42.20	42.86	42.50	44.78
LLL	33.30	38.06	42.50	48.27

Note: Columns correspond to test data and rows correspond to training data. The performance is measured by F-score (%).

6 Conclusions

In this paper, we described the challenges and practical issues with respect to the development of unified active learning framework for biomedical relation extraction. The proposed solution provides us with a very effective and practical way to design a robust and

economical method for biomedical relation extraction. Our active learning framework, not only achieves good performance with small amount of labelled data, but also provides us valuable savings in editorial time and maximal use of the labeling process. Experiments on PPI extraction have demonstrated the great potential and effectiveness of the proposed framework.

In the future, we plan to explore more examples selection methods, and apply this technique to extract more complex pathways in biomedical domain. Besides, in order to further explore the extension of our framework, we also plan to test active learning with other approaches, including Hidden Markov Model (HMM), Random Forest, Boosted Wrapper Induction.

References

[1] Faro A, Giordano D, Spampinato C. Combining literature text mining with microarray data: Advances for system biology modeling. *Brief Bioinform*, 2012, 13(1): 61-82.

[2] Hunter L, Cohen K. Biomedical language processing: What's beyond PubMed? *Mol Cell*, 2006, 21(5): 589-594.

[3] Huang M, Ding S, Wang H, Zhu X. Mining physical protein-protein interactions from the literature. *Genome Biology*, 2008, 9(Suppl 2): S12.

[4] Katrenko S, Adriaans P. Learning relations from biomedical corpora using dependency trees. In *Lecture Notes in Computer Science*, Tuyls K, Westra R, Saeys T *et al.* (eds.), Springer-Verlag, 2007, 4366, pp.61-80.

[5] Miwa M, Sætre R, Miyao Y, Tsujii J. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proc. the Conference on Empirical Methods in Natural Language Processing*, August 2009, pp.121-130.

[6] Yang Z, Lin H, Li Y. BioPPISVMExtractor: A protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. *Journal of Biomedical Informatics*, 2010, 43(1): 88-96.

[7] Li Y, Hu X, Lin H, Yang Z. Learning an enriched representation from unlabelled data for protein-protein interaction extraction. *BMC Bioinformatics*, 2010, 11(Suppl 2): S7.

[8] Landeghem S, Abeel T, Saeys Y, Peer Y. Discriminative and informative features for biomolecular text mining with ensemble feature selection. *Bioinformatics*, 2010, 26(18): 554-560.

[9] Bui Q, Katrenko S, Sloot P. A hybrid approach to extract protein-protein interactions. *Bioinformatics*, 2011, 27(2): 259-265.

[10] van Landeghem S, Saeys Y, Deu Baets B, van De Peer Y. Extracting protein-protein interactions from text using rich feature vectors and feature selection. In *Proc. the 3th International Symposium on Semantic Mining in Biomedicine*, September 2008, pp.77-84.

[11] Fayruzov T, De Cock M, Cornelis C, Hoste V. Linguistic feature analysis for protein interaction extraction. *BMC Bioinformatics*, 2009, 10: 374.

[12] Miyao Y, Sagae K, Sætre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 2009, 25(3): 394-400.

[13] Niu Y, Otasek D, Jurisica I. Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, 2010, 26(1): 111-119.

[14] Erkan G, Ozgur A, Radev D. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proc. the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, pp.228-237.

[15] Kim S, Yoon J, Yang J. Kernel approaches for genic interaction extraction. *Bioinformatics*, 2008, 24(1): 118-126.

[16] Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 2008, 9(Suppl 11): S2.

[17] Segura-Bedmar I, MartÍnez P, de Pablo-Sánchez C. Using a shallow linguistic kernel for drug-drug interaction extraction. *J. Biomed Inform*, 2011, 44(5): 789-804.

[18] Burr S. Active learning literature survey. Technical Report, University of Wisconsin-Madison. 2009.

[19] Dai H, Chang Y, Tsai, R T, Hsu W. New challenges for biological text-mining in the next decade. *J. Comput. Sci. Technol.*, 2010, 25(1): 169-179.

[20] Wang M, Hua X. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(2), Article No. 10.

[21] Long B, Chapelle O, Zhang Y, Chang Y, Zheng Z, Tseng B. Active learning for ranking through expected loss optimization. In *Proc. the 33rd Intarnational Conference on Research and Development in Information Retrieval*, July 2010, pp.267-274.

[22] He X. Laplacian regularized d-optimal design for active learning and its application to image retrieval. *IEEE Transactions on Image Processing*, 2010, 19(1):254-263.

[23] Bloodgood M, Callison-Burch C. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proc. the 48th Annual Meeting of the Association for Computational Linguistics*, July 2010, pp.854-864.

[24] Mohamed T, Carbonell J, Ganapathiraju M. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, 2010, 11(Suppl 1): S57.

[25] Klaus B. Incorporating diversity in active learning with support vector machines. In *Proc. the 20th International Conference on Machine Learning*, August 2003, pp.59-66.

[26] Huang M, Zhu X, Hao Y, Payan D, Qu K, Li M. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 2004, 20(18): 3604-3612.

[27] Wu F, Weld D. Open information extraction using wikipedia. In *Proc. the 48th ACL*, 2010, pp.118-127.

[28] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 2004, 5: 1205-1224.

[29] Riloff E. Automatically generating extraction patterns from untagged text. In *Proc. the 13th National Conference on Artificial Intelligence*, August 1996, pp.1044-1049.

[30] Quinlan J. Unknown attribute values in induction. In *Proc. the 6th Int. Workshop on Machine Learning*, June 1989, pp.164-168.

[31] Zhang H, Huang M, Zhu X. Protein-protein interaction extraction from bio-literature with compact features and data sampling strategy. In *Proc. the 4th BMEI*, October 2011, pp.1779-1783.

[32] Pyysalo S, Airola A, Heimonen J *et al.* Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 2008, 9(Suppl 3): S6.

Hong-Tao Zhang is a Ph.D. candidate in Department of Computer Science and Technology, Tsinghua University, China. His research interests include bioinformatics, text mining, and natural language processing.

Min-Lie Huang is a faculty member of Department of Computer Science and Technology, Tsinghua University. He received his Ph.D. degree from Tsinghua University in 2006. His research interests include text mining, natural language processing, opinion and review mining, and complex question answering. He has published papers in international conference proceedings such as ACL, AAAI, IJCAI, ICDM, COLING, NAACL, PAKDD and journals such as KAIS, Bioinformatics, JAMIA.

Xiao-Yan Zhu is a professor of Department of Computer Science and Technology, Tsinghua University. She got her bachelor degree at University of Science and Technology Beijing in 1982, master degree at Kobe University in 1987, and Ph.D. degree at Nagoya Institute of Technology, Japan, in 1990. She is teaching at Tsinghua University since 1993. She is the director of the State Key Lab of Intelligent Technology and Systems, Tsinghua-HP Joint Research Center, and Tsinghua-Waterloo Joint Research Center, Tsinghua University. She is the international research chair holder of IDRC, Canada, from 2009. She was the deputy head of Department of Computer Science and Technology, Tsinghua University from 2004∼2007. Her research interests include intelligent information processing, machine learning, natural language processing, query and answering system, and bioinformatics. She has authored more than 100 peer-reviewed articles in leading international conferences including SIGKDD, IJCAI, AAAI, ACL, ICDM, CIKM, COLING, and journals including Int. J. Medical Informatics, Bioinformatics, BMC Bioinformatics, Genome Biology and IEEE Trans. SMC.