Protein-Protein Interaction Extraction Based on Combinational Learning and Active Learning

21209163

2015.6

ώ ΄ ξ ώ

Dalian University of Technology

ŧ E ĥų r Ï ˬ سر ſ'

ŧ ŰŰ ſ Ï Χ̨ ˋY م غي Ï

Ғ˜ U ˋl ƚ γˬ φ ĭ Ï Ӣ Ғ˜ U ˋl ƚ

r ˋl ƚ ą ſ ŧ ' n ŧ n Ã

شج سر Χ̨ Ã ʜ ʜ ſ

Ғ Ă Ï ˬ Ĩ o ſ

r E_____

ŧ o E_____ E_____ ____ ____

Protein-Protein Interaction, PPI

PPI

Q-statistic

AIMed 71% F 92.9% AUC

PPI

5

LLL 4 AUC AIMed

BioInfer

Protein-Protein Interaction Extraction Based on Combinational Learning and Active Learning

Abstract

The continuous development of life science and technology has resulted in an explosive increase of the biomedical literature. Therefore, intelligent methods that can automatically extract useful information from massive literatures are urgently needed. As the Internet is becoming increasingly perfecting, information extraction technology has rapidly developed. And, it has a significant impact on biomedical research. Protein-Protein Interaction (PPI) extraction is an important application of information extraction technology in biomedical field. It is aiming at the potential knowledge mining from the molecular level. Centering on the PPI extraction problem, this paper carries out the following researches.

To solve the features deficiency and the problem of limited decision-making ability brought by a single classifier, this paper proposes a combinational learning method. This approach is focusing on the features design and multi-classifier integration. In feature selection, depending on the sentence contexts and the syntactic structures, rich features are extracted to construct feature vectors besides, the information gain method is used to screen the feature ; in classifier integration, three classifiers with different decision-making schemes and higher precision are chosen and used separately, including Support Vector Machine, Maximum Entropy and Naive Bayes. Then, we integrate the classification results by adopting linear weighted method to make sure that the classifier which performed better has a larger weight. The combinational learning approach gained 71% of the F-score and 92.9% of the AUC-score on AIMed corpus.

The combinational learning approach is only fit for the situation with enough labeled corpus. However, there is relatively few labeled corpus in practice. Thus, in order to solve this problem, this paper proposes an active learning on the basis of combinational learning. This method is based on the selection approach of the uncertain samples. It repeatedly selects the samples with most useful information from large amounts of unlabeled corpus and ignores the useless ones. Then, annotate the selected samples and add them to the original training set. The active learning method can not only achieve a better PPI extraction performance, but also reduce the hand-annotated work. The experiments of active learning also achieve higher AUC-scores on the other four corpuora except for LLL this approach shows a better generalization performance in large corpora as AIMed and BioInfer.

Key Words Protein-Protein Interaction Extraction; Rich Feature; Feature Selection Combinational Learning; Active Learning

1

1.1

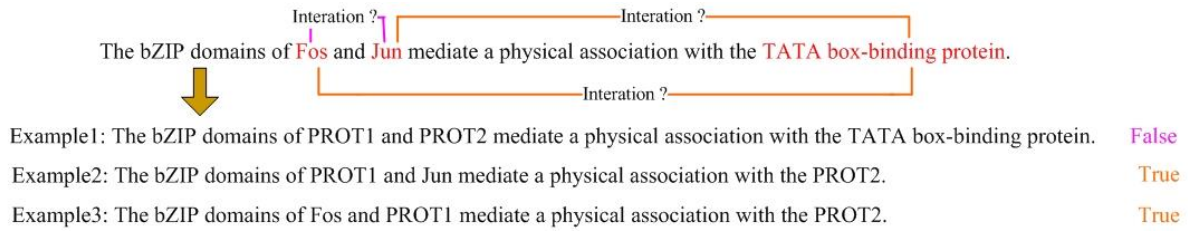
Information extraction, IE

- Protein- Protein Interaction, PPI

PPI

“ The bZIP domains of Fos and Jun mediate a physical association with the TATA box-binding protein. ”

PPI 1.1



1.1

Fig. 1.1 Example of protein-protein interaction extraction

1.1 3 Fos, Jun, TATA box-binding protein 3

3

PROT1 PROT2 1.1

Example 1 Fos Jun False Example 2 Example 3 True

PPI

PPI

PPI

AIMed[1] BioInfer[2] IEPA[3] HPRD50[4] LLL[5]

PPI

1.2

PPI

PPI

1. 2.1 ɲ ǐ

[6]

1. 2. 2 ɲ ʮ

Part-of-speech Tagging

[7]

1. 2. 3 л ф й

[8]

PPI

PPI

PPI

1

[9]

“ ” Niu [10] Liu [11]

AIMed 53.5% 54.% F

2

[12]

Walk-weighted Subsequence Kernel All-paths Graph Kernels Tree Kernel Hash Graph Kernel Airola [13]

PPI

Kim [14]

PPI AIMed F 56.6% Qian [15] PPI

5

Zhang [16]

AIMed

F AUC 60.2% 85.3% Li [17]

5

PPI

PPI

Miwa [18]

Yang [19]

AIMed F AUC 64.41%

88.46% Li [20]

AIMed F

AUC 69.40% 92.00% Li [21]

4

Bui

[22] AIMed

F 61.2% Qian [23]

AIMed 63.1% F

1.3 ✗

PPI

1 PPI

Q-statistic

2

GDep

Stanford Parser Enju

1.4

5

1 PPI

2

3

5

4 PPI

PPI

2 ĩ ŭ ĩ

2.1 Ḫ ŭ

[24]

Information Retrieval, IR

[25]

1

2

3

1

2

3

2.2 ꞵ ⅎ

2.2.1 ʂ ʁ ㄱ

[26]

GDep Stanford Parser

GDep[27] Tsujii

LR

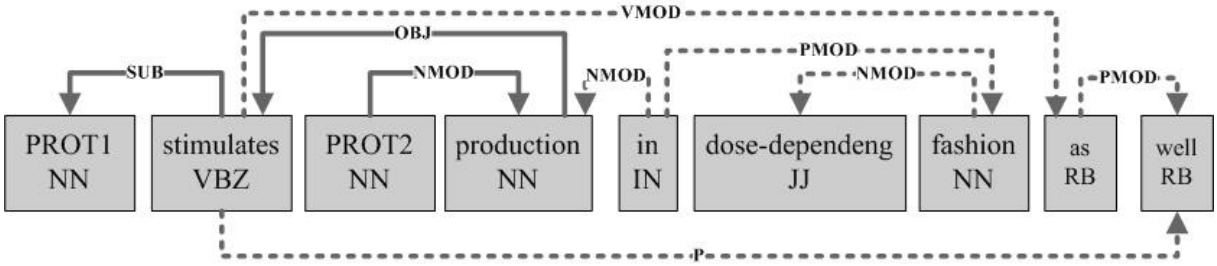
2.1 GDep “ PROT1 stimulates PROT2 production in dose-dependent fashion as well. ”

“ stimulates ” 2.1

“ stimulates ” “ as ”

“ VMOD ” “ stimulates ” “ as ”

“ stimulates ” “ as ”



2.1 GDep

Fig. 2.1 Example of the output produced by Gdep

Stanford Parser [28]

Stanford Parser

Stanford Parser

POS

2.2 2.3 Stanford Parser “ PROT1 stimulates PROT2 production in dose-dependent fashion as well. ”

2.2

2.3

```
(ROOT
  (NP
    (NP (CD PROT1) (NNS stimulates))
    (NP
      (NP (JJ PROT2) (NN production))
      (PP (IN in)
        (NP
          (NP (JJ dose-dependent) (NN fashion))
          (CONJP (RB as) (RB well))
          (NP (NNP #)))))))
```

2.2 Stanford Parser

Fig. 2.2 The phrase structure produced by Stanford Parser

num(stimulates-2, PROT1-1)
root(ROOT-0, stimulates-2)
amod(production-4, PROT2-3)
dep(stimulates-2, production-4)
amod(fashion-7, dose-dependent-6)
prep_in(production-4, fashion-7)
prep_in(production-4, #-10)
conj_and(fashion-7, #-10)

2.3 Stanford Parser

Fig. 2.3 The dependency representation produced by Stanford Parser

2.2.2 ㅂ ㅓ

Enju[29] Tsujii

Enju

Enju

Phrase Structures Predicate -Argument Structures

2.4 " PROT1 stimulates PROT2 production in dose-dependent fashion as well. "

12

1 2

3 4 5 6

7 8 9

10 11 12

```
ROOT    ROOT    ROOT    ROOT    -1  ROOT    ROOT    stimulates  stimulate   VBZ VB  1
stimulates  stimulate   VBZ VB  1   verb_arg12  ARG1    PROT1   prot-NUMBER-    CD  CD  0
stimulates  stimulate   VBZ VB  1   verb_arg12  ARG2    production  production  NN  NN  3
PROT2   prot-NUMBER-    CD  CD  2   adj_arg1    ARG1    production  production  NN  NN  3
dose-dependent  dose-dependent  JJ  JJ  5   adj_arg1    ARG1    fashion fashion NN  NN  6
in  in  IN  IN  4   prep_arg12  ARG1    production  production  NN  NN  3
in  in  IN  IN  4   prep_arg12  ARG2    fashion fashion NN  NN  6
well    well    RB  RB  8   adj_arg1    ARG1    stimulates  stimulate   VBZ VB  1
as  as  RB  RB  7   adj_arg1    ARG1    well    well    RB  RB  8
ROOT    ROOT    ROOT    ROOT    -1  ROOT    ROOT    #   -SHARP- #   -SHARP- 9
#   -SHARP- #   -SHARP- 9   noun_arg1   ARG1    UNKNOWN UNKNOWN UNKNOWN UNKNOWN -1
```

2.4 Enju

Fig. 2.4 The predicate argument structure produced by Enju

2.3

[30]

1

IG [31]

2

(DF)[32] t, t

DF

DF DF

2.4 U й

[33]

[34,35]

1

SVM 20 90 Vapnik

SVM VC

SVM

SVM

[36]

SVM

SVM

2.1 ~ 2.4

Linear Function

$$K(x,y)=(x\ \ y) \tag{2.1}$$

Polynomial Function

$$K(x,y)=(x\ \ y)\quad c+ \tag{2.2}$$

Radial Basis Function, RBF

$$K(x,y)=\exp\left(-\frac{|x-y|^2}{2\sigma^2}\right) \tag{2.3}$$

Sigmoid Sigmoid Function

$$K(x,y)=\tanh\left(v(x,y)\ +\right) \tag{2.4}$$

2

Maximum Entropy, ME E.T.Jaynes 1957

[37]

3

[38]

[35]

2.5 Ї ̈ й

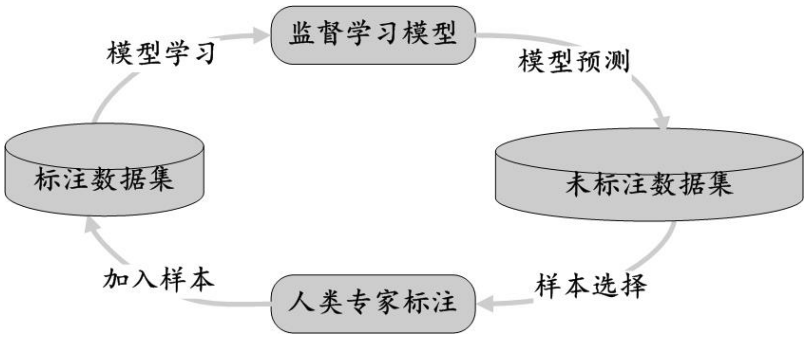
Active Learning

[39]

2.5

[40]

“ ”



2.5

Fig. 2.5 Framework of active learning

2.6 ۴

2.6.1

P Precision R Recall F F-score, F AUC AUC-score, AUC

0 1

1

2.5 2.6

$$P = \frac{TP}{TP + FP} \tag{2.5}$$

$$R = \frac{TP}{TP + FN} \tag{2.6}$$

TP True Positive ; FP False Positive ; FN False Negative

F F 2.7

$$F = \frac{2 \times P \times R}{P + R} \tag{2.7}$$

AUC Area Under the Curve

[41] AUC ROC Receiver Operating Characteristic Curve

AUC 0.5 1 1

AUC 2.8

$$AUC = \frac{\sum r_i - n_+(n_+ + 1)/2}{n_+ \times n_-} \tag{2.8}$$

n_+ n_- r_i i

2.6.2

AIMed BioInfer HPRD50 IEPA LLL

5

XML

True False 5

AIMed 225 PubMed DIP

200

Biolnfer
Biolnfer 5

HPRD50 HRPD Human Protein Reference Database 50
ProMiner

IEPA 200 PubMed

LLL LLL05 Learning Language in Logic 2005

2.1 5 PPI
5 AIMed
5 PPI
HPRD50 8921 LLL 27 5
LLL 1 1 AIMed

2.1 5 PPI
Tab. 2.1 Statistics of the five PPI corpora

AIMed	225	1955	5834	1000	4834
BioInfer	836	1100	8921	2437	6484
HPRD50	50	145	433	163	270
IEPA	200	486	817	335	482
LLL	45	77	330	164	166

З л U й ĩ ŭ

PPI

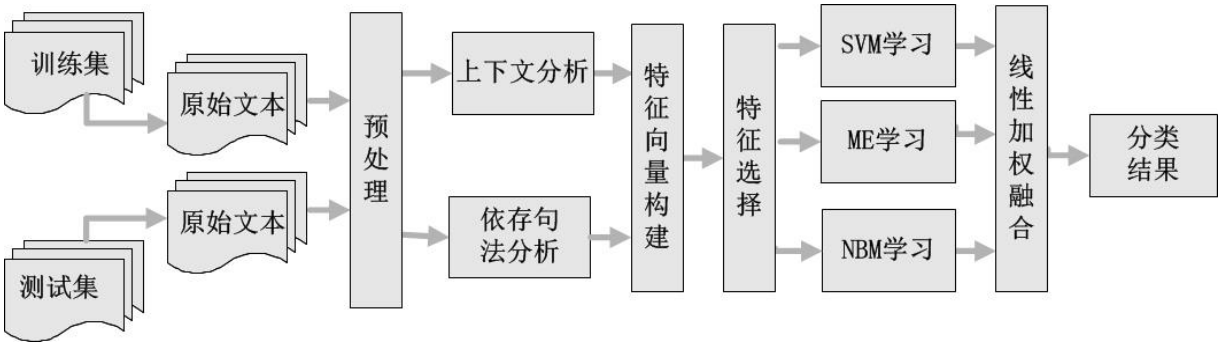
PPI

PPI

Q-statistic

3.1

Stanford Parser GDep Enju



3.1

Fig. 3.1 Framework of PPI extraction based on combination of learners

3.1

PPI 5 XML

a the

“ . ” “ - ” “ ” “ < ” “ % ”

n n^2 2

C_n^2

“ PROT1 ” “ PROT2 ”

3.2 ϥ

PPI Tikk [42]

PPI

GDep Stanford Parser Enju

GDep

Stanford Parser

Enju

3.2.1 ϱ Ґ

1 N

N

N

N 2 3 N

N N

2

N

“ PROT1 ”

“ PROT1_Left_ N ” “ PROT 1_Right_ N

, PROT 2_Left_ N ” “ PROT2 ” “ PROT 2_ Right _

N ”

3

N

“ PROT1 ” “ PROT2 ”

“ PROT1 ” “ PROT2 ”

4

5

6

5 5 5

7

“ neither, hardly, not, nor, never ”

3.2.2 [8]

1

2

3

4

Stanford Parser

Stanford Parser

5

3.2

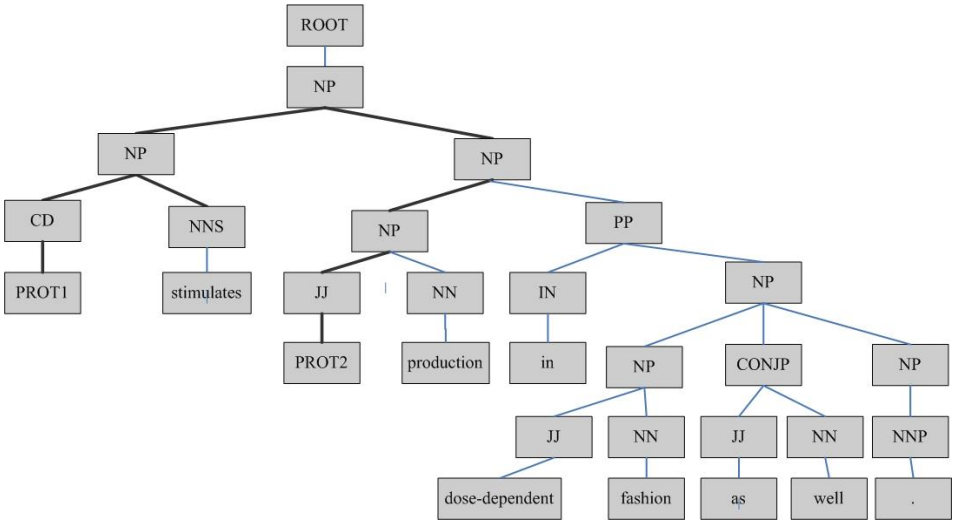
“ PROT1 stimulates PROT2 production in dose-dependent fashion as well. ” Stanford Parser

“ PROT1 ” “ PROT2 ”

“ PROT1 ” “ PROT2 ”

(1),(2),(3),(4),(5) Stanford Parser

Stanford Parser



3.2

Fig. 3.2 Structure path between two proteins

6

“ PROT1 ” “ PROT2 ”

GDep

PROT1 PROT2

Brown VB VBZ VBD

0

2.1

PROT1 PROT2

1

7 walk

“ PROT1 ” “ PROT2 ”

“ ” …

Kim [14] e-walk v-walk

“ ” “ ”

(6) (7) GDep GDep

8

Zhang [43] Shortest Path-enclosed Tree, SPT 5

Enju

“ …

” “ ” “ ”

9

(8) (9) Enju Enju

3.3

3.1

$$IG(t) = H(C) - H(C|t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum p(c_i|\bar{t})\log p(c_i|\bar{t}) \quad (3.1)$$

$p(c_i)$ c_i $p(t)$ t

$p(c_i|t)$ t c_i $p(\bar{t})$

t $p\left(c_i|\bar{t}\right)$ t c_i m

3.4 ฯ ฝ U

PPI

0.5

Q-statistic[44]

Q_i Q_j

$$Q_{ij} = \frac{p^{11}(i,j)\,p^{00}(i,j) - p^{01}(i,j)\,p^{10}(i,j)}{p^{11}(i,j)\,p^{00}(i,j) + p^{01}(i,j)\,p^{10}(i,j)} \qquad (3.2)$$

$p^{11}(i,j)$ Q_i Q_j

$p^{10}(i,j)$ Q_i Q_j

$p^{01}(i,j)$ Q_i Q_j

$p^{00}(i,j)$ Q_i Q_j

Q_{ij} -1 1 $Q_{ij} = 0$

L Q_{ij}

3.3

$$\overline{Q} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} Q_{ij} \tag{3.3}$$

SVM

3.2 3.3

0.04 0.35

0.29

0.226 0

Q-statistic

PPI

$$\begin{cases} K(x, x') = \sum_{n=1}^{N} s_n K_n(x, x') \\ \sum_{n=1}^{N} s_n = 1, \quad s_n \geq 0, \quad n \end{cases} \tag{3.4}$$

N s_n n

0.1

0.3 0.6

3.5

5 k

10 10-fold Document-level Cross-validation

10 [45] 10

9

F AUC

10

10 F AUC

3.5.1 AIMed

3.2.1 7 3.2.2 9

AIMed 30 3.1

3.1 N AIMed

1000

29 F AUC 3.3 3.4

3.1 3.3 3.4 10000 F AUC

10000 F

AUC

10000 F AUC 0.655 0.908

300000 F AUC 10.1 5.1

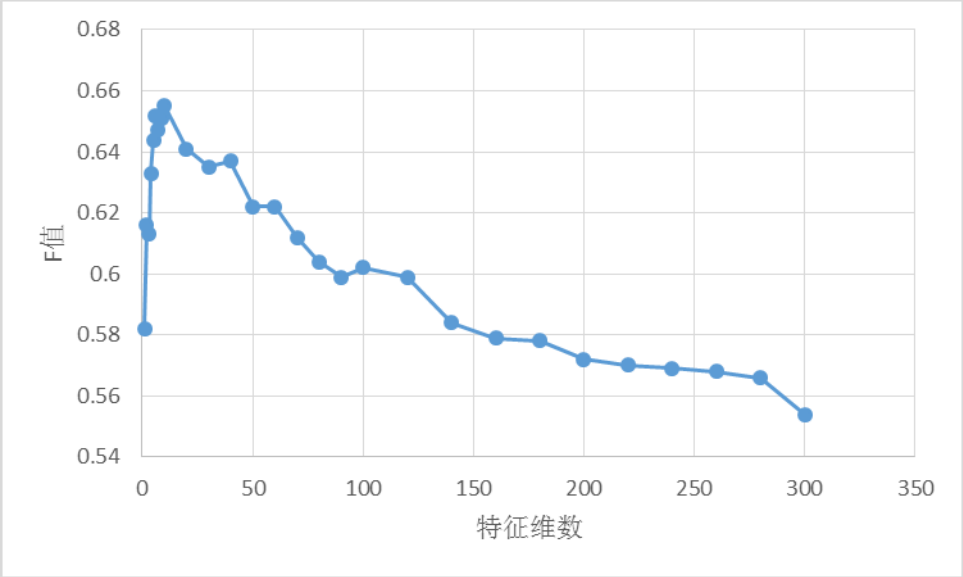
3.1 AIMed

Tab. 3.1 Experiment results of different dimension features on AIMed

	(×1000)	P	R	F	AUC
1	1	0.587	0.581	0.582	0.866
2	2	0.599	0.642	0.616	0.887
3	3	0.587	0.647	0.613	0.891
4	4	0.592	0.689	0.633	0.901
5	5	0.596	0.708	0.644	0.904
6	6	0.600	0.723	0.652	0.905
7	7	0.591	0.721	0.647	0.906
8	8	0.588	0.741	0.652	0.906
9	9	0.586	0.741	0.651	0.907
10	10	0.586	0.749	0.655	0.908
11	20	0.555	0.766	0.641	0.908
12	30	0.542	0.773	0.635	0.906
13	40	0.547	0.771	0.637	0.906
14	50	0.525	0.771	0.622	0.902
15	60	0.519	0.785	0.622	0.900
16	70	0.509	0.780	0.612	0.897
17	80	0.498	0.778	0.604	0.893
18	90	0.491	0.781	0.599	0.892
19	100	0.493	0.789	0.602	0.887
20	120	0.485	0.790	0.599	0.886
21	140	0.473	0.780	0.584	0.880
22	160	0.465	0.783	0.579	0.878
23	180	0.462	0.788	0.578	0.876
24	200	0.458	0.781	0.572	0.870
25	220	0.454	0.780	0.570	0.870
26	240	0.453	0.781	0.569	0.869
27	260	0.451	0.781	0.568	0.868
28	280	0.449	0.784	0.566	0. 868
29	300	0.436	0.775	0.554	0.857

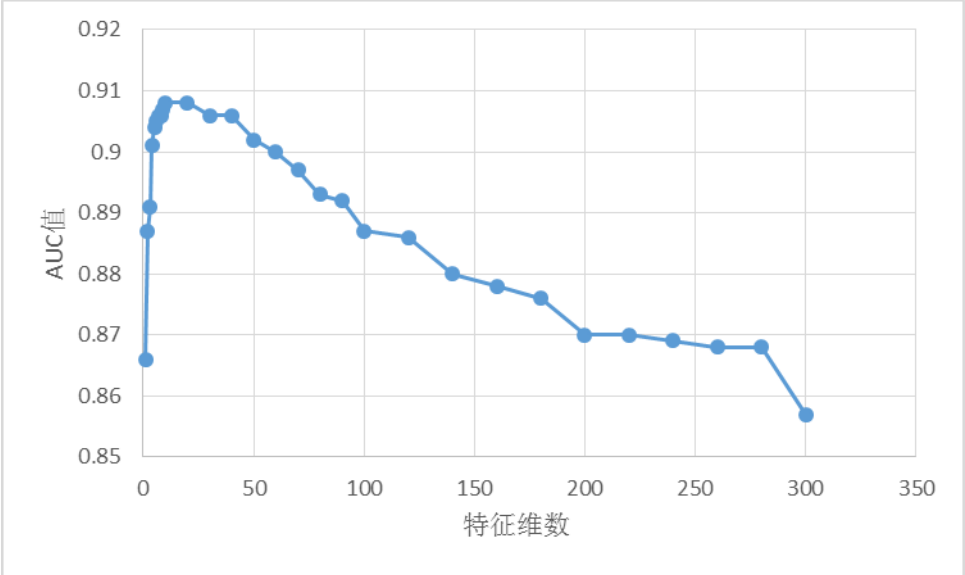
10000 300000

10000



3.3 AIMed F

Fig. 3.3 Results of F-score between different dimension features on AIMed



3.4 AIMed AUC

Fig. 3.4 Results of AUC-score between different dimension features on AIMed

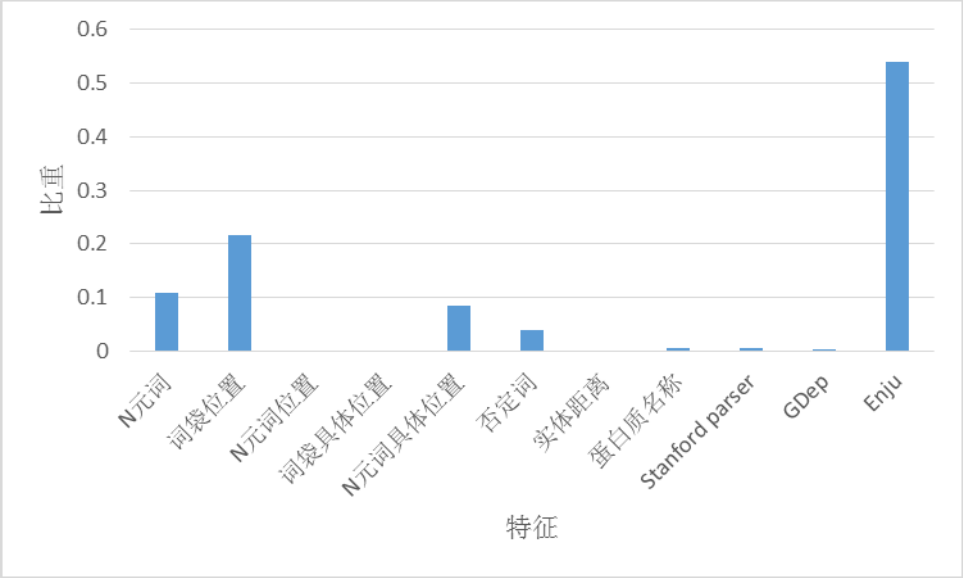
10000 3.5

3.5 Enju

50% GDep Stanford parser Enju

N

N



3.5 10000

Fig. 3.5 The proportion of different features 10,000-dimensional feature

3.5.2 U ﻱ ﺱ AIMed p

3.2

3.2

AIMed F 60% AUC 90%

F AUC

F 0.69 AUC 0.92

6:3:1 F AUC

0.71 0.929

3.2 AIMed

Tab. 3.2 Fusion results on AIMed corpus

ME:SVM:NBM	P	R	F	AUC
1:0:0	0.56	0.84	0.67	0.913
0:1:0	0.59	0.73	0.65	0.908
0:0:1	0.46	0.89	0.61	0.9
1:1:1	0.69	0.69	0.69	0.920
6:3:1	0.66	0.77	0.71	0.929

3.5.3 F n سر AIMed p

3.3 PPI PPI AIMed

3.3

Airola Miwa

Bui

Yang

3.3 F AUC

3.3 AIMed

Tab. 3.3 Performance comparison between different methods on AIMed

	P	R	F	AUC
Airola et al. [13]	0.53	0.62	0.56	0.85
Miwa et al. [18]	0.59	0.66	0.62	0.88
Bui et al. [22]	0.55	0.68	0.61	—
Yang et al. [19]	0.58	0.71	0.64	0.89
	0.66	**0.77**	**0.71**	**0.929**

3.6

PPI

Q-statistic

AIMed F AUC 71% 92.9%
PPI 5 LLL 4
AUC PPI

4 л ӏ ̄ й ї̃ Ӧ́

3 PPI

PPI PPI

PPI

SVM

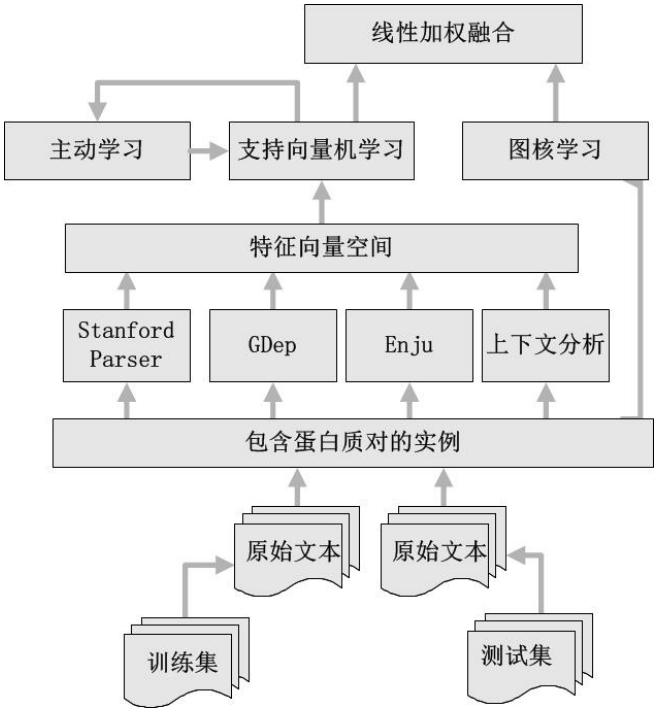
4.1 PPI

PPI PPI

1

2

3 3.3



4.1

Fig. 4.1 Framework of PPI extraction based on active learning

4.1 Ι ̄ й

PPI

PPI

4.1.1 ɲ ⊦

[46]

Logistic

5

4 SVM

0.1

4.2

基于不确定样本选择的主动学习算法

Input:

从已标注语料中划分出来的训练集: O_1;
剩余的四个语料作为未标注语料集: O_2;
从已标注语料中划分出来的测试集: T;
使用特征向量的SVM分类器模型: S;
使用特征向量的ME分类器模型: M;
每次选择的新样本的数量: N。

Output:

用模型S训练T得到的准确率P，召回率R，F值以及AUC值。

1 训练O_1，得到模型S和M。
2 **Repeat**
3 使用最大熵模型M对测试集O_2进行分类，并从测试集O_2中选择出模型M最不确定的N个新样本。
4 对选出来的N个新样本进行标注，并添加到O_1中。
5 将此N个样本从O_2中去除掉。
6 重新训练O_1，得到新的分类器模型S和M。
7 使用当前的模型S预测测试集T。
8 计算P，R，F值以及AUC值。
9 **Until** 没有模型M最不确定的样本。
10 **Print** 最终的准确率P，召回率R，F值以及AUC值。

4.2

Fig. 4.2 Active learning algorithm based on sample selection uncertainty

4.1.2 Л †

[47]

Seung[48] Freund[49] Query By Committee, QBC QBC

4.1.3 Л . †

[50]

4.2 ´ U

PPI

4.2.1 Л Ч

PPI

PPI

PPI

1 2

PPI svm[light][51]

4.2.2 ЭЭ

Airor [13]

PSS

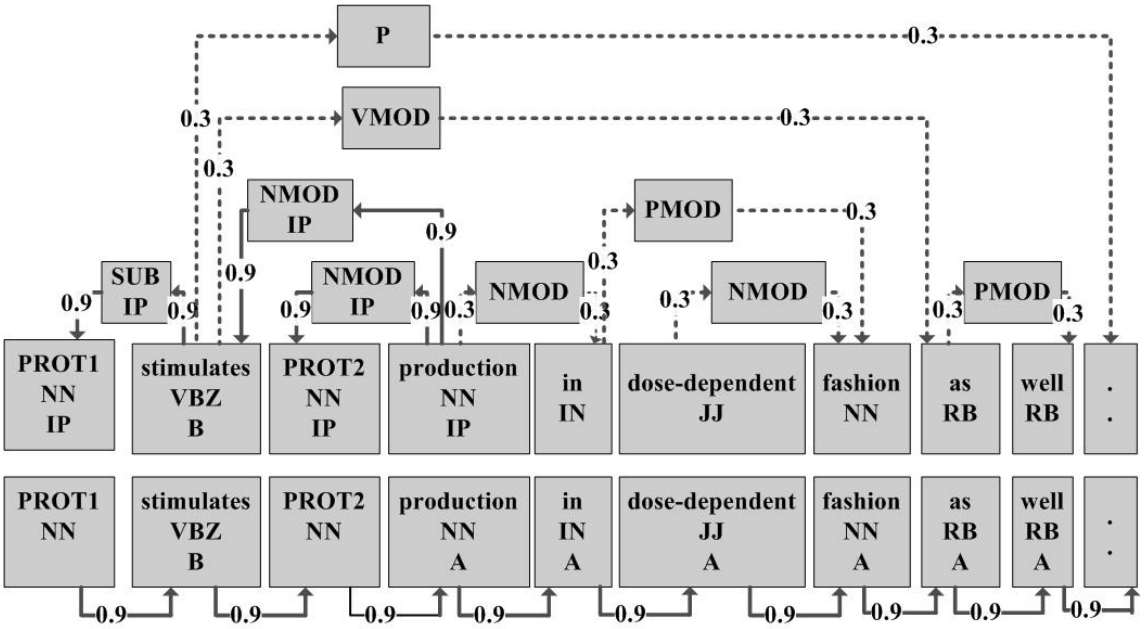
LOS PSS

IP

IP

PSS LOS

Airor PSS LOS 0.9
0.3 4.3 PSS LOS “ PROT1 stimulates PROT2 production in dose-dependent fashion as well. ”



4.3

Fig. 4.3 Two directed subgraph of graph kernel

L

A **G** 4.1

$$G = L^T \sum_{i=1}^{\infty} A^n L \tag{4.1}$$

L $N \times L$ j i
L_{ij} 1 0 **A** $L \times L$ i j
A_{ij} 0.9 0.3
0 4.2 **G** $\mathbf{G}'$
$K(\mathbf{G}, \mathbf{G}')$

$$K(\mathbf{G}, \mathbf{G}') = \sum_{i=1}^{L} \sum_{j}^{L} G_{ij} G'_{ij} \tag{4.2}$$

4.2.3 U

PPI

3

3.1

0.7 0.3

0 1

0.1

4.3 ғ ⊣

4.3.1 AIMed p

AIMed

PPI PPI AIMed

Stanford Parser GDep Enju

PPI

Stanford GDep

Enju

4.2

F AUC

4.2

GDep F AUC 8.5 5.5

PPI Stanford F

Enju F 65.3% AUC 90.3%

4.2 AIMed

Tab. 4.2 The experiment results after different procedures on AIMed corpora

	P	R	F	AUC
	0.559	0.535	0.541	0.829
+GDep	0.622	0.636	0.626	0.884
+Stanford	0.618	0.652	0.632	0.885
+Enju	0.649	0.665	0.653	0.903
+	0.647	0.688	0.662	0.905
+	0.673	0.663	0.664	0.909

GDep Enju

F 10.6 AUC 7.3

PPI

AIMed

F /66.4% AUC /90.9%

4.3 PPI AIMed

F AUC

4.3 AIMed F

66.4% AUC 90.9% Miwa [18]

60.8% F 86.8% AUC Kim [14]

F 56.6%

F AUC Yang [19]

4

71.1% Qian [15]

Qian [23]

F 63.1%

4.3 AIMed

Tab. 4.3 Performance comparison between different methods on AIMed corpora

	P	R	F	AUC
Miwa [18]	0.550	0.688	0.608	0.868
Kim [14]	0.614	0.533	0.566	-
Yang [19]	0.577	0.711	0.644	0.885
Qian [15]	0.591	0.576	0.581	0.833
Qian [23]	0.651	0.613	0.631	-
	0.673	0.663	**0.664**	**0.909**

4.3.2 F n سر 5 γ p

5 4.4

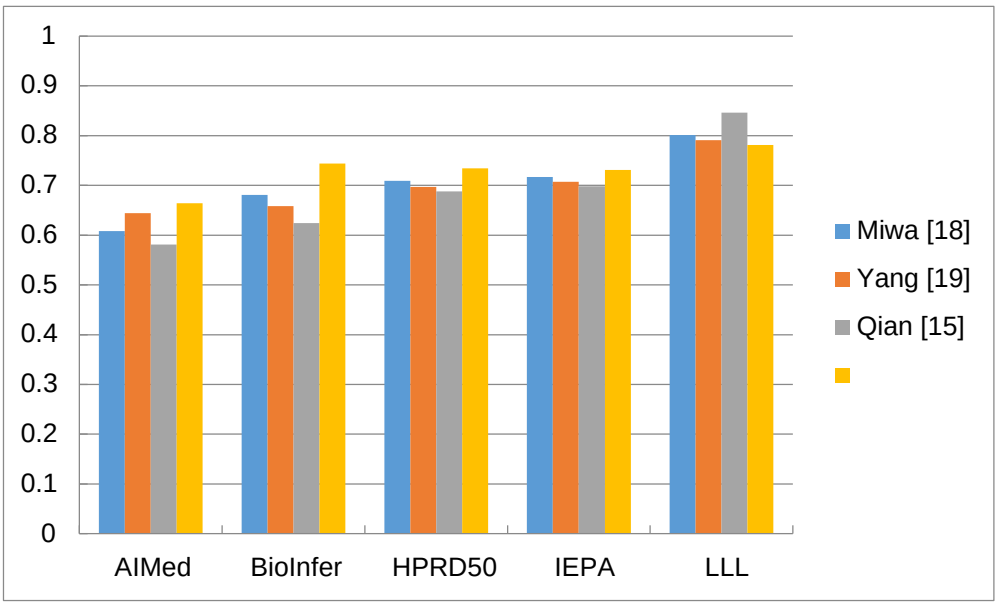
4.4 5

Tab. 4.4 Performance comparison between different methods on five corpora

		AIMed	BioInfer	HPRD50	IEPA	LLL
Miwa [18]	F	0.608	0.681	0.709	0.717	0.801
	AUC	0.868	0.859	0.822	0.844	0.863
Yang [19]	F	0.644	0.658	0.697	0.707	0.791
	AUC	0.885	0.850	0.744	0.757	0.830
Qian [15]	F	0.581	0.624	0.688	0.698	0.846
	AUC	0.833	0.836	0.837	0.828	0.899
	F	0.664	0.744	0.734	0.731	0.781
	AUC	0.909	0.921	0.845	0.855	0.866

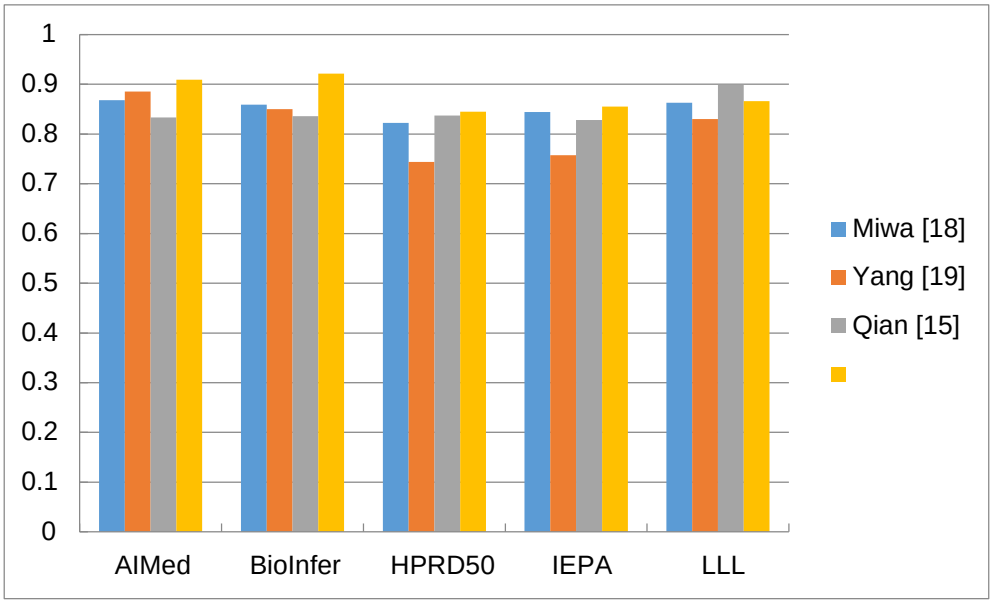
4.4 4.5

F AUC



4.4 F 5

Fig. 4.4 Comparison results of F-score between different methods on five corpora



4.5 AUC 5

Fig. 4.5 Comparison results of AUC-score between different methods on five corpora

4.4 4.5 4 AIMed BioInfer HPRD50 IEPA F AUC BioInfer F AUC

LLL

5 PPI

[52,53]

5 4 4.5 F 4.6 AUC

4.5 F

Tab. 4.5 Cross-corpus results comparison measured with F-score

	AIMed		BioInfer		HPRD50		IEPA		LLL		Avg.rank
	F	Rank	F	Rank	F	Rank	F	Rank	F	Rank	
AIMed	-	-	0.530	2	0.674	2	0.672	2	0.734	2	2
BioInfer	0.485	1	-	-	0.685	1	0.740	1	0.753	1	1
HPRD50	0.425	2	0.471	4	-	-	0.640	4	0.706	4	3.5
IEPA	0.385	3	0.548	1	0.643	4	-	-	0.708	3	2.75
LLL	0.345	4	0.491	3	0.650	3	0.646	3	-	-	3.25

4.6 AUC

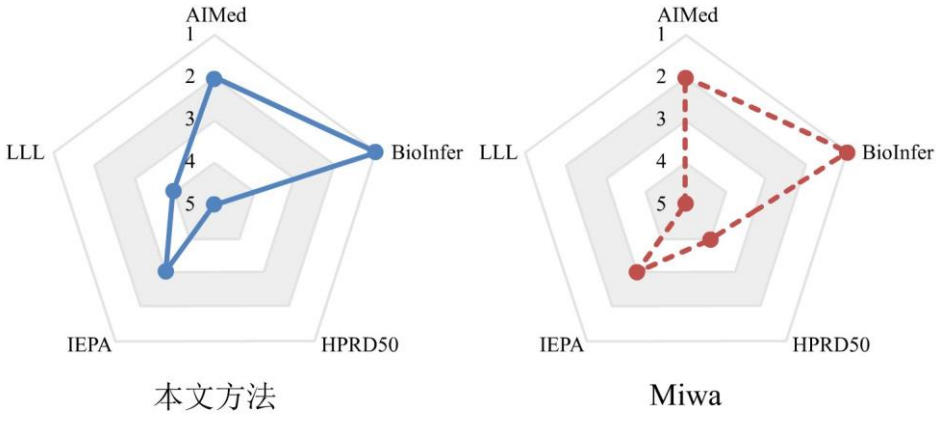
Tab. 4.6 Cross-corpus results comparison measured with AUC score

	AIMed		BioInfer		HPRD50		IEPA		LLL		Avg.rank
	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank	
AIMed	-	-	0.702	2	0.804	1	0.758	2	0.782	4	2.25
BioInfer	0.792	1	-	-	0.801	2	0.841	1	0.841	1	1.25
HPRD50	0.714	2	0.652	4	-	-	0.725	3	0.810	3	3
IEPA	0.668	3	0.721	1	0.764	3	-	-	0.827	2	2.25
LLL	0.610	4	0.658	3	0.752	4	0.723	4	-	-	3.75

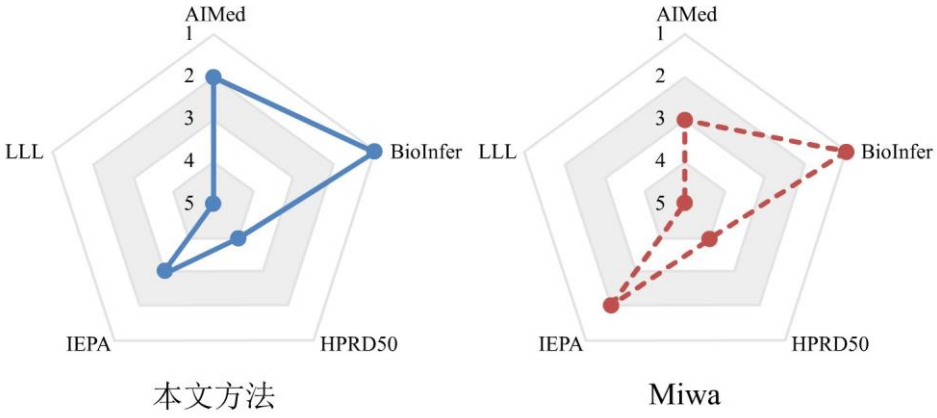
Rank F AUC
4.5
AIMed BioInfer F Rank 1 LLL
F Rank 4 Avg.rank Rank
Avg.rank PPI

4.5 AIMed F
48.5% BioInfer AIMed F 66.4%

F AUC
BioInfer IEPA
BioInfer IEPA
4.6 AUC 61.0% 84.1%
AUC
AUC 4.4 AUC 84.5% 92.1%
4.5 4.6 F AUC Avg.rank
4.6 4.7 Miwa [13]
Avg.rank 4.6
4.7
Avg.rank



4.6 5 F Avg.rank
Fig.4.6 Comparison results between different Avg.rank values of F-score on five corpus



4.7 5 AUC Avg.rank

Fig.4.7 Comparison results between different Avg.rank values of AUC-score on five corpus

4.6 4.7 Miwa [13]

BioInfer

HPRD50 LLL Avg.rank

PPI

4.4

Gdep Enju

PPI

1

Q-statistic

PPI

10

AIMed 4 P R

F AUC PPI

2

5 PPI

10

LLL

4 AUC AIMed BioInfer

1

2

参考文献

[1] Bunescu R, Ge R, Kate R J, et al. Comparative experiments on learning information extractors for proteins and their interactions [J]. Artificial intelligence in medicine, 2005, 33(2): 139-155.

[2] Pyysalo S, Ginter F, Heimonen J, et al. BioInfer: a corpus for information extraction in the biomedical domain [J]. BMC bioinformatics, 2007, 8(1): 50.

[3] Ding J, Berleant D, Nettleton D, et al. Mining MEDLINE: abstracts, sentences, or phrases? [C] // Pacific Symposium on Biocomputing. 2002, 7: 326-337.

[4] Fundel K, Küffner R, Zimmer R. RelEx—Relation extraction using dependency parse trees [J]. Bioinformatics, 2007, 23(3): 365-371.

[5] Nédellec C. Learning language in logic-genic interaction extraction challenge [C] // Proceedings of the 4th Learning Language in Logic Workshop (LLL05), 2005, 7.

[6] Bunescu R, Mooney R, Ramani A, et al. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline [C] //Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. Association for Computational Linguistics, 2006: 49-56.

[7] Huang M, Zhu X, Hao Y, et al. Discovering patterns to extract protein-protein interactions from full biomedical texts [J]. Bioinformatics, 2004, 20(18): 3604-3612.

[8] Shiguihara-Juárez P N, de Andrade Lopes A. Learning trees for extraction of protein-protein interaction [C]//Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing-Volume 2. Springer-Verlag, 2013: 347-358.

[9] Mitsumori T, Murata M, Fukuda Y, et al. Extracting protein-protein interaction information from biomedical text with SVM [J]. IEICE Transactions on Information and Systems, 2006, 89(8): 2464-2466.

[10] 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 [J]. Bioinformatics, 2010, 26(1): 111-119.

[11] Liu B, Qian L, Wang H, et al. Dependency-driven feature-based learning for extracting protein-protein interactions from biomedical text [C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 757-765.

[12] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge university press, 2000.

[13] Airola A, Pyysalo S, Björne J, et al. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning [J]. BMC bioinformatics, 2008, 9(Suppl 11): S2.

[14] Kim S, Yoon J, Yang J, et al. Walk-weighted subsequence kernels for protein-protein interaction extraction [J]. BMC bioinformatics, 2010, 11(1): 107.

[15] Qian L, Zhou G. Tree kernel-based protein-protein interaction extraction from biomedical literature [J]. Journal of biomedical informatics, 2012, 45(3): 535-543.

[16] Zhang Y, Lin H, Yang Z, et al. Neighborhood hash graph kernel for protein-protein interaction extraction [J]. Journal of biomedical informatics, 2011, 44(6): 1086-1092.

[17] Li L, Guo R, Jiang Z, et al. An Approach to Improve Kernel-Based Protein-Protein Interaction Extraction by Learning from Large-Scale Network Data [J]. Methods, 2015.

[18] Miwa M, Sætre R, Miyao Y, et al. Protein-protein interaction extraction by leveraging multiple kernels and parsers [J]. International journal of medical informatics, 2009, 78(12): e39-e46.

[19] Yang Z, Tang N, Zhang X, et al. Multiple kernel learning in protein-protein interaction extraction from biomedical literature [J]. Artificial intelligence in medicine, 2011, 51(3): 163-173.

[20] Li L, Zhang P, Zheng T, et al. Integrating Semantic Information into Multiple Kernels for Protein-Protein Interaction Extraction from Biomedical Literatures [J]. PloS one, 2014, 9(3): e91898.

[21] Li L, Jin L, Zheng J, et al. The Protein-Protein Interaction extraction based on full texts [C]//Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on. IEEE, 2014: 493-496.

[22] Bui Q C, Katrenko S, Sloot P M A. A hybrid approach to extract protein–protein interactions [J]. Bioinformatics, 2011, 27(2): 259-265.

[23] Qian W, Fu C, Cheng H. Semi-supervised method for Extraction of Protein-Protein Interactions using hybrid model [C]//Proceedings of the 2013 Third International Conference on Intelligent System Design and Engineering Applications. IEEE Computer Society, 2013: 1268-1271.

[24] , , . [J]. , 2009, 44(21): 166-170.

[25] . [D]. : ,2008.

[26] , . [J]. ,2009,11(2):100-112.

[27] Miwa M, Pyysalo S, Hara T, et al. A comparative study of syntactic parsers for event extraction [C]//Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2010: 37-45.

[28] , . Stanford Berkeley [J]. ,2013,9(8):1985-1986.

[29] Miyao Y, Sagae K, Saetre R, et al. Evaluating contributions of natural language parsers to protein-protein interaction extraction[J]. Bioinformatics, 2009, 25(3): 394-400.

[30] , , , . [J]. , 2012, 27(2): 161-166.

[31] , . [J]. , 2012, 48(27): 119-122.

[32] . [D]. , 2010.

[33] . [D]. : ,2009.

[34] Bartlett P L, Traskin M. Adaboost is consistent [J]. Journal of Machine Learning Research, 2007, 8: 2347-2368.

[35] Sun Y, Todorovic S, Li J, et al. Unifying the error-correcting and output-code adaboost within the margin framework [C]//Proceedings of the 22nd international conference on Machine learning. ACM, 2005: 872-879.

[36] . [M]. : 2012.

[37] Berger AL, Pietra VJD, Pietra SAD. A maximum entropy approach to natural language processing [J]. Computational linguistics, 1996, 22(1): 39-71.

[38] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine learning, 1997, 29(2-3): 103-130.

[39] , , . [J]. , 2011, 37(8): 954-962.

[40] . [D]. : ,2013.

[41] . [D]. : ,2012.

[42] Tikk D, Solt I, Thomas P, et al. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction[J]. BMC bioinformatics, 2013, 14(1): 12.

[43] Zhang M, Zhang J, Su J, et al. A composite kernel to extract relations between entities with both flat and structured features [C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 825-832.

[44] Yule G U. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c[J]. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1900: 257-319.

[45] Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature [C]//EACL, 2006, 18: 401-408.

[46] , , , . [J]. ,2008, 45( ): 300-304

[47] , , . [J]. ,2010,46(13): 11-14.

[48] Seung H S, Opper M, Sompolinsky H. Query by committee [C]//Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992: 287-294.

[49] Freund Y, Seung H S, Shamir E, et al. Selective sampling using the query by committee algorithm [J]. Machine learning, 1997, 28(2-3): 133-168.

[50] , , . [J]. ,2013, 26(12):1121-1129.

[51] Joachims T. SVM light support vector machine [EB/OL]. 2002. http://svmlight.joachims.org .

[52] Van Landeghem S, Saeys Y, De Baets B, et al. Extracting protein-protein interactions from text using rich feature vectors and feature selection [C]//3rd International symposium on Semantic Mining in Biomedicine (SMBM 2008). Turku Centre for Computer Sciences (TUCS), 2008: 77-84.

[53] Zhang Y, Lin H, Yang Z, et al. Hash subgraph pairwise kernel for protein-protein interaction extraction [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2012, 9(4): 1190-1202.

r ǐ "

1 . .
2015 11
2 WANG J, **LIU M**, LIN H, et al. Combining Active Learning and Composite Kernel for Protein-protein Interaction Extraction[J]. Journal of Computational Information Systems, 2015, 11(8): 2823-2832.
3 . .
2014SR156358

r ə ö

˛ ˋHʜ Ĩ r ˛ Ï ѕr r

ɤ ˛ Л Ï ѽ м Ṣ ɾ

Ḧ Ϥ ᵩ Ĩ ˬ ˩ ł м Ï˝ II

r ˋH ɜ ˋY ĩ Ĩ Ï˝ II ˩ ɻ

˩ ɻ ┼ Ḧ м r ɾ

r E_____

ɤ o E_____ E_____ ____ ____

o E_____ E_____ ____ ____