生物医学文献蛋白质关系抽取——从实体识别到网络构建*

杨志豪，赵哲焕，李彦鹏，胡运翠，谢 冬，林鸿飞

大连理工大学 计算机科学与技术学院，大连 116024

E-mail: yangzh@dlut.edu.cn

摘 要：本文介绍了一个从实体识别到PPI网络构建的生物医学文献蛋白质关系抽取系统。该系统采用特征耦合泛化策略进行蛋白质实体识别；采用基于扩展语义相似度的方法进行蛋白质名均一化；融合了基于特征的核、树核以及图核进行蛋白质关系抽取；并实现了蛋白质关系网络的可视化。该系统在DIP数据库的一个子集上获得59.88%的综合分类率（F-score），取得了优于其他系统的性能。

关键词：文本挖掘；信息抽取；蛋白质关系抽取；支撑向量机；多核学习

A Protein-protein Interaction Extraction System from Medical Literature—From NER to Network Construction

Yang Zhihao, Zhao Zhehuan, Li Yanpeng, Hu Yuncui, Xie Dong, Lin Hongfei

Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024

E-mail: yangzh@dlut.edu.cn

Abstract: This paper presents a system to automatically extract protein-protein interactions from biomedical literature and construct the PPI network. The system applies Feature Coupling Generalization to tag protein names, uses extended semantic similarity to normalize protein mentions, combines feature-based kernel, tree kernel and graph kernel to extract PPI, and finally visualize the PPI network. Experimental evaluations show that our system can achieve state-of-the-art performance with respect to comparable evaluations, with 59.88 % F-score on a DIP subset.

Keywords: text mining; information extraction; protein-protein interaction; support vector machines; multiple kernels learning

1 引言

生物医学文献中的蛋白质关系抽取可以用于蛋白质知识网络的建立、蛋白质关系的预测以及新药的研制等，对于研究生物过程有着重要意义。当前，已经建立了许多结构化存储蛋白质关系的数据库，如 MINT、BIND、DIP。然而，随着生物医学文献数量的迅速增长，很难依靠人工挖掘蛋白质关系信息。因此，生物医学文献中自动抽取蛋白质关系信息成为当前非常重要的研究课题。

近年来，越来越多机器学习的方法用于蛋白质关系信息抽取。机器学习的方法中，基于核的方法是一种特征抽取的有效方法。它保持对象的原始表达形式，通过计算一对实体的核函数的值使用这些对象。许多核方法包括子序列核、树核、最短路径核以及图核已被用于蛋白质关系信息抽取。但是，每种核方法都只是利用了句子的部分结构信息来计算相似度。本文提出了一个生物医学文献蛋白质关系抽取系统。该系统采用特征耦合泛化策略进行具有较高精度的蛋白质实体识别；基于扩展语义相似度的方法的蛋白质名均一化；融合了基于特征的核、树核以及图核进行蛋白质关系抽取。

2 方法

整个系统由蛋白质实体识别、蛋白质均一化、蛋白质关系抽取以及蛋白质关系网络可视化部

* 本文承国家自然科学基金（60673039，61070098）；国家“八六三”基金（2006AA01Z151）；中央高校基本科研业务费专项资金资助（DUT10JS09）；辽宁省博士启动基金（20091015）资助。

分组成。系统的输入为 Pubmed 生物医学文摘，输出为可视化的蛋白质关系网络。系统的实现要借助 Entrez 基因数据库和 GO（Gene Ontology）本体。

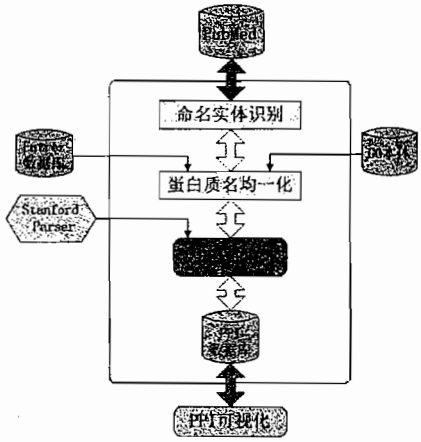

图 1　基于多核学习的生物医学文献蛋白质关系抽取系统框架图

2.1　蛋白质实体识别

本文使用特征耦合泛化策略进行具有较高精度的蛋白质实体识别[1]。特征耦合泛化的基本思想是从实例区分特征和类别区分特征。实例区分特征和类别区分特征分别表示指向实例和类别的强指示器。直觉上，它们在大量未标注语料中的共现会携带一些信息，而由于数据的稀疏性，在已标注语料中有可能得不到这些信息。在一个 17GB 大小的用于基因命名实体识别任务的未标注语料上，用此方法得到进一步优化的实体名称特征效果比精心设计的词汇特征好很多。利用经过本方法优化后的词典，在 BioCreative2 测试集上做的向前最大匹配方法得到的 F 值为 86.2。我将这个词典和基于条件随机域的基因标注标签结合得到了 89.05 的 F 值。

2.2　蛋白质名均一化

蛋白质名均一化任务的目的是为生物医学文献中提到的蛋白质名确定其在数据库中对应的唯一的标识符，从而建立起这些实体与生物学数据库之间的联系。与命名实体识别相比，均一化任务面临着语义方面的挑战，因为它需要对蛋白质所表达的真实含义进行检测，还需要给出蛋白质的标准形式。本文使用了一种基于扩展语义相似度的蛋白质名均一化方法。该方法利用 MEDLINE 摘要信息和基因本体描述信息，为数据库中的蛋白质标识符生成了扩展的语义信息；然后通过比较歧义蛋白质名的上下文信息和其不同语义描述信息之间的相似性，为歧义蛋白质名确定能够表达真实含义的唯一蛋白质标识符。该方法使用 BioCreative II 蛋白质均一化任务的语料进行测试，F 值达到了 81.2%。

2.3　蛋白质关系抽取

核方法可以看做是计算一对对象的相似度的函数，不同的核从两个句子的不同的方面计算相似度。融合多种相似度可以避免丢失重要的特征，从而生成一个有用的相似度度量。在本系统中，融合了包括基于特征的核、树核以及图核用于蛋白质关系抽取。

2.3.1　基于特征的核

在本系统中，基于特征的核方法主要使用了以下三种特征：词特征、蛋白质实体名的距离特征和关键词特征。

2.3.2 树核

卷积树核的目的在于从子结构上获取有用的结构信息。卷积树核 $K_C(T_1, T_2)$（'C'指卷积)是一种特殊的卷积核，通过计算两个句法分析树 T_1 和 T_2 的相同子树结构的数目作为二者的语义相似度。其中 N_j 是树 T_j 中的节点集，$\Delta(n_1, n_2)$ 使用递归的算法计算以 n_1 和 n_2 为根的相同的子树结构数目。

$$K_C(T_1, T_2) = \sum_{n_1 \in N_1, n_2 \in N_2} \Delta(n_1, n_2) \tag{1}$$

（1）语法分析树核

通常选取最短路径依存树(SPT)作为树核计算范围，SPT 是语法分析树中连接两个实体的最短路径上附着的子树。但在某些情况下，SPT 中的信息不足以判定两个实体间的关系。比如，在句子"ENTITY1 and ENTITY2 interact with each other"中，"interact"可以帮助判断"ENTITY1"和"ENTITY2"间存在交互关系。然而，SPT(图 2 中的虚点圆)中包含的信息不足以判定它们的关系。我们使用了一个简单的启发式规则来扩展 SPT。默认情况下，我们采用 SPT 作为计算范围。当 SPT 中叶节点的个数少于四的时候，则对 SPT 扩展，使它包含 SPT 之外的重要文本信息，如图 2 所示(实线圆部分)。

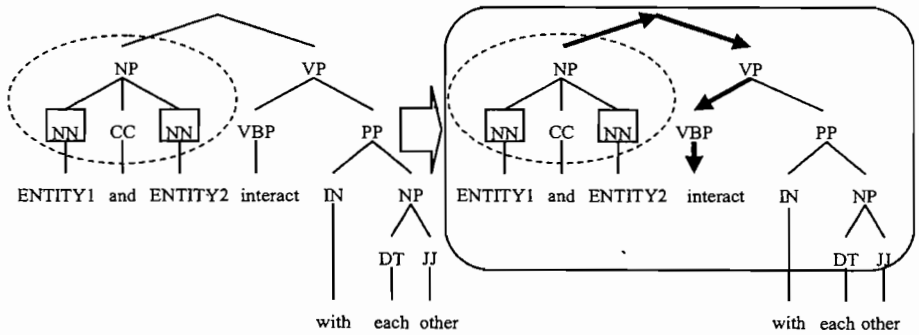

图 2　最短路径依存树扩展

（2）依存路径树核

另一种树核使用的树的结构信息来源于语法分析器依存分析的输出结果。与 SPT 分析树相似，在某些情况下，依存路径树也需要扩展。以"The expression of rsfA is under the control of both ENTITY1 and ENTITY2."为例，ENTITY1 和 ENTITY2 之间的路径树是"(DEPENDENCY (CONJ (ENTITY1, ENTITY2)).", 这棵路径树中的信息不足以判定这两个实体之间的关系。我们的解决办法是，当路径长度小于三时，扩展依存路径的长度。上例中 ENTITY1 和 ENTITY2 之间的路径可以扩展为"(DEPENDENCY (PREP(control, of)) POBJ((of, ENTITY1)) (CONJ(ENTITY1, ENTITY2)))"。

2.3.3 图核

图核通过比较目的关系之间的相同顶点实现两输入图的相似度计算。我们使用的图核是由 Airola 等提出的全路径图核[2]。

2.3.4 核的融合

这些核方法从不同方面计算了两个句子的相似度，融合这些相似度可以避免遗漏重要的特征。为了实现基于不同分析结构的核的融合，我们对多个核 K_m 的均一化结果求和，其中 m 代表了核的个数：

$$K(x, x') = \sum_{m=1}^{M} \sigma_m K_m(x, x') \tag{2}$$

$$\sum_{m=1}^{M} \sigma_m = 1,\ \sigma_m \geqslant 0,\ \forall m \tag{3}$$

3 实验结果和讨论

3.1 多核方法的 PPI 抽取性能

我们使用 AImed 语料进行了多核方法的 PPI 抽取方法性能测试。该语料具有较大的规模，近年来被看作是蛋白质关系抽取方法的评测标准。我们对语料集分别进行文档级的十倍交叉验证实验。采用的性能评测指标是当前 PPI 抽取系统主要使用的 F 值。此外，我们还使用了 AUC 值来评测结果。其优点在于它不受数据集的类别的分布的影响，目前已被作为新的性能评测标准。

表 1 不同核方法在 AImed 语料上的性能

方法	P	R	F	AUC
基于特征的核	46.32	61.1	52.69	80.71
树核	43.71	64.65	52.24	79.19
图核	52.66	64.56	57.20	83.27
基于特征的核+树核	50.44	68.49	58.05	84.19
基于特征的核+图核	51.33	69.58	59.02	84.68
树核+图核	53.43	68.57	59.66	85.51
基于特征的核+树核+图核	57.4	70.75	63.9	87. 83

表 1 是不同核方法在 AImed 语料上的结果。其中，图核的性能最好。当融合图核和树核时，F 值提高了 2.46 个百分点，AUC 值提高了 2.24 个百分点。当进一步与基于特征的核融合，取得了最好的性能：F 值 63.9%，AUC 值 87.83%。实验结果表明，三种核方法的融合可以取得非常好的性能改进。表 2 是我们的方法与其他方法的实验结果对比。表现最好的系统融合了多层语义信息，文献[3]通过多种语法分析器的多核合并实现蛋白质关系抽取，F 值达到了 63.5%，AUC 值达到 87.9%。而我们的方法并未使用多种语法分析器，也获得了与其接近的性能。

表 2 不同方法在 AImed 语料上的性能比较

方法	P	R	F	AUC
本文方法	57.4	70.75	63.9	87. 83
Miwa 等[3]	60.4	69.3	63.5	87.9
Miyao 等[4]	54.9	65.5	59.5	
Airola 等[2]	52.9	61.8	56.4	84.8

3.2 PPI 抽取系统的综合性能

3.1 节描述的是 PPI 抽取阶段的性能。而作为完整的医学文献蛋白质关系抽取系统，命名实体识别和均一化阶段的性能也会对整个系统的抽取性能造成影响。在实验中用到的训练集是由五个公共语料 Aimed、Bioinfer、HPRD50、IEPA 和 LLL 构成。使用的测试集与 BioRAT[5]和 IntEx[6]中用到的测试集相同，包含 394 个关系。在表 3 中的召回率与表 2 中数据（用来测试分类的性能）不同，是抽取到的关系数与 394（关系的总数）的比率。

BioRAT 和 IntEx 系统中未经过实体均一化处理过程，这意味着无法获得蛋白质对应的唯一标识符，因而无法进行有效集成。表 3 第一行对应的是同样未经过标注化处理的本文方法实现系统 MKBioPPIExtractor 的性能。与其他系统相比，其获得了最好的综合分类率（F-score）（59.88%）。表 3 第二行对应的是经过标注化处理的 MKBioPPIExtractor 系统性能。可以看到其性能，特别是召回率有所下降。原因在于：经过标注化处理，许多蛋白质名由于表达不规范，并未被转换成对应

的唯一标识符，导致提取出的蛋白质交互对减少，造成召回率下降。即使如此，其各项指标依然高于 BioRAT 和 IntEx 系统。

表 3　不同方法在 DIP 数据库子集上的综合抽取性能比较

	P	R	F
MKBioPPIExtractor（未均一化）	53.88	68.74	59.88
MKBioPPIExtractor（均一化）	29.95	66.53	41.3
IntEx	26.94	65.66	38.20
BioRAT	20.31	55.07	29.68

4　可视化

本系统在关系网络数据模型与可视化方面主要利用了 JUNG 工具包（http://jung.sourceforge.net/）。JUNG 提供了一组通用性强并且易扩展的编程接口，用于对图或网络的数据结构进行建模、分析和可视化。图 3 所示为 krogan 蛋白质关系网络数据集中的子图。

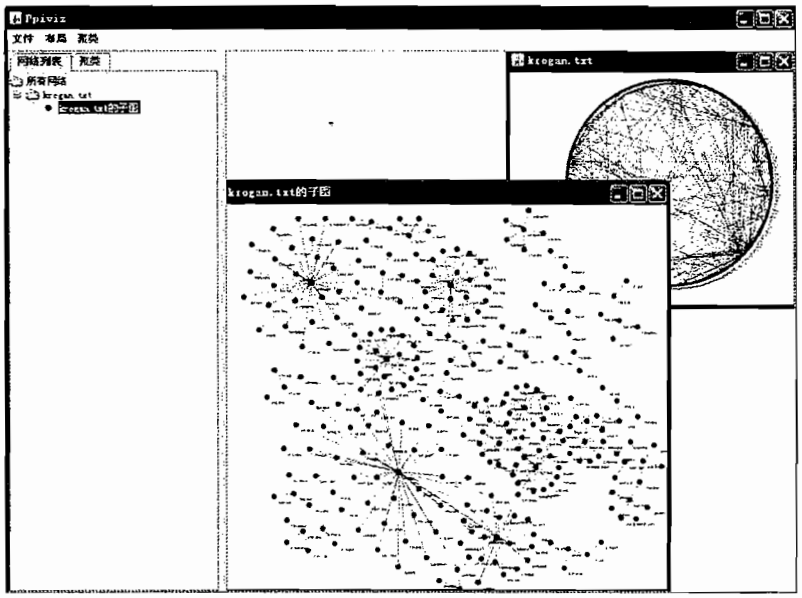

图 3　krogan 蛋白质关系网络

5　总结

本文介绍了一个从实体识别到 PPI 网络构建的生物医学文献蛋白质关系抽取系统 MKBioPPIExtractor。该系统采用特征耦合泛化策略进行具有较高精度的蛋白质实体识别；采用基于扩展语义相似度的方法进行蛋白质名均一化；融合了基于特征的核、树核以及图核进行蛋白质关系抽取；并实现了蛋白质关系网络的可视化。

参 考 文 献

[1] Yanpeng Li, Hongfei Lin, Zhihao Yang Incorporating rich background knowledge for gene named entity classification and recognition BMC Bioinformatics 2009, 10: 223.

[2] 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.

[3] Miwa M, Sætre R, Miyao Y, Ohta T, Tsujii J. Combining Multiple Layers of Syntactic Information for Protein-Protein Interaction Extraction. In: Salakoski T, Rebholz-Schuhmann D, Pyysalo S, editors., Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine, Turku, Finland. 2008. p.101-108.

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

[5] Comey DP, Buxton BF, Langdon WB, Jones DT. BioRAT: extracting biological information from full-length papers. Bioinformatics 2004; 20(17): 3206-3213.

[6] Ahmed ST, Chidambaram D, Davulcu H, Baral C. IntEx: a syntactic role driven protein - protein interaction extractor for bio-medical text. In: Proc. the ACLISMB Workshop on Linking Biological Literature, Ontologies.