依存信息在蛋白质关系抽取中的作用*

刘兵 徐华 钱龙华 周国栋

(苏州大学计算机科学与技术学院，江苏，苏州，215006)

(江苏省计算机信息处理技术重点实验室，江苏，苏州，215006)

E-mail: liubingnlp@gmail.com

摘 要：基于核函数的蛋白质关系（PPI）抽取可以捕获结构化句法信息，取得了较高的性能，但计算复杂度过高。本文结合词汇、句法信息，重点探讨了依存信息对基于特征向量的蛋白质关系(PPI)抽取的影响。研究表明，依存信息和基本短语块信息可以有效提高基于特征向量的PPI抽取性能。本文在多个PPI语料库上进行了实验，其中在AIMed语料上的实验取得了F测度为54.7的较好性能，是目前基于特征向量的PPI抽取系统的最好水平。

关键字：蛋白质关系，支持向量机，依存信息

The Role of Dependency Information for Protein-Protein Interaction Extraction

Liu Bing, Xu Hua, Qian Longhua, Zhou Guodong

(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu, 215006)

(Jiangsu Key Lab. of Information Processing Technology, Suzhou, Jiangsu, 215006)

E-mail: liubingnlp@gmail.com

Abstract: Kernel-based PPI extraction systems achieve promising performance because of their capability to capture structural syntactic information, but at the expense of computational complexity. This paper investigates the combination of diverse lexical, syntactic and especially dependency information in feature-based protein-protein interaction extraction using SVM. Our study illustrates that dependency information as well as base phrase chunking information is very effective for feature-based PPI extraction. Additionally, we execute our feature-based method on multiple PPI corpora and the experimental evaluation on the AIMed corpus shows that our system achieves a promising performance of 54.7 in F-measure, surpassing other state-of-the-art feature-based ones.

Keywords: PPI, SVM, Dependency Information

1. 引言

蛋白质作为最主要的生命活动载体和功能执行者，其复杂多样的结构功能、相互作用和动态变化能在分子、细胞和生物体等多个层次上全面揭示生命现象，对生命活动过程中蛋白质相互作用关系（PPI）的研究有助于揭示生命过程的许多本质问题，因而PPI抽取成为生物医学领域信息抽取的重点研究方向，是生物医学文本挖掘的主要任务之一，具有重要的研究意义。

目前，计算机辅助的自动文本挖掘技术大量使用，可以快速地获取医学信息并将其构建为计算机可操作的知识库，方便数据分析和挖掘。这些方法大致可分为三类：基于规则的方法，基于

*基金资助：国家自然科学基金(60873150，60970056，90920004)。

作者简介：刘兵（1986-），男，硕士研究生，主要研究方向：信息抽取；徐华（1988-），男，硕士研究生，主要研究方向：信息抽取；钱龙华（1966-），男，副教授，硕士生导师，研究方向：自然语言处理；周国栋（1967-），男，教授，博士生导师，研究方向：自然语言处理。

特征向量的方法和基于核函数的方法。

基于规则的方法采用一些预先定义的词或短语模式规则匹配可能出现的PPI关系。然而，预定义的规则不可能包含所有的PPI关系模式，并且由于规则的领域适应性，当在新的领域使用基于规则的方法时，这些规则都需要修正。这种方法可以获得较高的准确率，但召回率非常低。

基于特征向量的方法近年来得到广泛应用。在以往的研究中，Mitsumori等[1]抽取了蛋白质实体附近的词特征，探索了词汇特征效果。Sugiyama等[2]从包含PPI的句子中抽取了动词和名词信息，进一步研究了词汇信息尤其是动词对PPI抽取的作用。此外，Giuliano等[3]还探讨了lemma等浅层语言学信息。然而，上述系统都没有考虑任何句法和依存信息，而研究表明这些信息在新闻领域的关系抽取中有很好的效果[4]。因此，Sætre等[5]将句法信息，浅层依存关系信息和词汇特征结合起来进行PPI抽取，显著提高了PPI抽取的性能。但是目前对于依存信息的研究并不深入，所采用的特征也不能有效捕获依存树中的结构化信息，系统性能相对于新闻领域仍有很大差距。

当前机器学习领域的另一个热门课题就是核函数的研究和使用。基于核函数的方法直接以结构树为处理对象，再使用支持核函数的分类器进行关系抽取。然而受制于计算复杂度，该方法往往不能应用于实际的PPI抽取系统中，这也促使我们考虑在基于特征向量的方法中对依存信息做进一步探索，以最大化结构化信息尤其是依存信息的作用。

本文第2部分介绍了基于特征向量的PPI抽取方法，分析了基准系统所用的各种特征。第3部分详细描述了依存信息驱动的PPI抽取系统所用的特征。第4部分给出了数据处理方案、实验流程和结果分析。最后一部分是本文的结论和展望。

2. 基于特征向量的 PPI 抽取

对于基于特征向量的信息抽取方法来说，PPI 抽取可以看作是一个分类问题。首先，系统要将标注好的 PPI 实例构造成一个特征集合，并映射到一个 n 维的特征向量空间；然后，在特征向量空间上运用机器学习方法，这个过程可以分为两个阶段：在训练时，分类学习算法利用标注好的 PPI 实例学习得到一个分类器；测试时，利用该分类器判断待测试的关系实例所属的关系类别，预测 PPI 是否存在。

以往的信息抽取研究表明，浅层句法信息在关系抽取中的作用非常明显。因此，我们先提取词汇，基本短语块和简单句法树信息构建一个基准系统，作为与加入依存信息的 PPI 抽取系统的对比。基准系统用到的特征如下：

2.1 词汇特征

本系统用到四种类型的词汇信息：1）蛋白质实体的名称；2）两个蛋白质实体之间的词；3）第一个蛋白质实体之前的词；4）第二个蛋白质实体之后的词。为避免引入噪音，我们只考虑第一个蛋白质前和第二个蛋白质后的两个词。

2.2 基本短语块特征

基本短语块特征是用 Sabine Buchholz 的 perl 脚本[1]从完全句法树中获得的，而句法树则是由

1 http://ilk.kub.nl/~sabine/chunklink/

Stanford Parser[2]生成。类似词汇特征，我们提取了以下基本短语块特征：

- CPHBNULL：实体之间没有短语块
- CPHBFL：实体之间仅有一个短语块时，该短语块的核心词
- CPHBF：实体之间至少两个短语块时，其第一个短语块的核心词
- CPHBL：实体之间至少两个短语块时，其最后一个短语块的核心词
- CPHBO：实体之间除去首尾两个的其他短语块的核心词
- CPP：连接两个实体所在短语块的短语块类型。

为防止以上基本短语块特征过于具体而造成数据稀疏问题，我们构造了一系列组合特征。这些特征是将上述基本短语块特征（CPP 除外）与它们对应的短语块类型结合起来得到的。

2.3 句法树特征

- PTP：句法树中两个实体之间的路径（经过去重处理）。

3. 依存信息驱动的 PPI 抽取

依存树可以揭示句子中的长距离依存信息，且能避免非结构化特征中出现的噪音，可以为关系抽取提供更为有效的信息。目前，利用依存信息进行 PPI 抽取的研究主要集中于基于核函数的方法，比如 Airola 等[6]采用全依存路径图核，Kim 等[7]采用加权路径子串核分别进行了实验，并获得了不错的性能。除此之外，Sætre 等[5]， Miyao 等[8]和 Miwa 等[9][10]采用复合核方法将平面特征与结构化信息结合起来进行 PPI 抽取，大幅度提高了系统性能。虽然目前基于核函数的方法在性能方面要比基于特征向量的高，但是核函数方法都存在计算复杂性的瓶颈问题，而另一方面，依存信息在基于特征向量的 PPI 抽取中的作用还有待深入研究。所以在本节我们将抽取一系列依存特征，考察它们在 PPI 抽取中的表现。

依存树特征也是借助 Stanford Parser 得到的。Stanford Parser 依存分析的输出格式是：依存类别（word1, word2），其中 word1 是核心词，word2 依赖于该核心词，依存类别则由 Stanford Parser 预先定义。根据这些依存关系对，我们可以构建一个句子的依存树，并且抽取它的如下特征，记作 DependenecySet1：

- DP1TR：依存树中蛋白质 PROT1 到根节点的路径
- DP2TR：依存树中蛋白质 PROT2 到根节点的路径
- DP12DT：依存树中两个蛋白质之间的依存关系类别
- DP12：连接两个蛋白质路径上的词和依存类型的组合
- DP12S：DP12 中的每个单词及其依存类型的组合
- DPFLAG：判断两个蛋白质是否具有直接依存关系

以句子“PROT1 contains a sequence motif binds to PROT2.”（记为句 1）为例， Stanford Parser 生成的语法关系及构造的依存树如下：

2 http://nlp.stanford.edu/software/lex-parser.shtml

nsubj(contains-2,PROT1-1)
det(motif-5, a-3)
nn(motif-5, sequence-4)
nsubj(binds-6, motif-5)
ccomp(contains-2, binds-6)
prep_to(binds-6, PROT2-8)

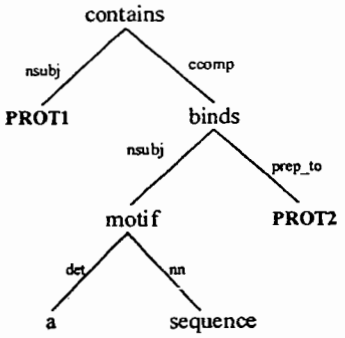

图 1 句 1 的语法关系及其依存树

除了上述依存特征，我们也探索了依存树中动词对 PPI 抽取的影响。不同于新闻领域，在生物医学领域文本的关系抽取中，动词扮演着更为重要的角色，这是因为主要动词的变化可以很容易导致 PPI 关系极性的改变。而以往的研究中并没有对这一问题给予足够关注，因此本文抽取了以下动词及其在依存树中的位置特征，记作 DependencySet2：

- FVW：DP12 特征中位于第一个蛋白质之前的动词
- LVW：DP12 特征中位于第二个蛋白质之后的动词
- MVW：DP12 特征中的其他动词
- #FVW: FVW 中动词的数目
- #LVW: LVW 中动词的数目
- #MVW: MVW 中动词的数目

4. 实验结果与分析

4.1 实验设置

我们采用 AIMed 语料库作为主要实验数据集，AIMed 是一种广泛应用于 PPI 抽取领域的语料库，它包含 225 篇从 MEDLINE 中提取的文章摘要。另外，我们也在其他四个经常使用的 PPI 语料库[3]上进行了实验。

实验使用的所有实例都是由至少含有两个蛋白质实体的句子生成的，实例的数目由句子中蛋白质的个数 n 决定，根据组合原理，每个句子可以产生 $\binom{n}{2}$ 个实例。同该领域的其他研究一样，我们不考虑这些关系实例的方向和类型，预处理时去除了语料库中的 59 个自相关的 PPI 实例，保留了 154 个嵌套的 PPI 实例，最终本系统抽取了 1002 个关系正例和 4794 个关系负例，正负例比例在公认的正常范围内。

实验中，我们选择 SVM 作为分类器。SVM 分类器本质上是二元分类器，所以它非常适合判断 PPI 是否存在的任务。在本系统中，我们使用了 Joachims 等开发的二元分类工具 SVMLight[4]。

实验设置方面，我们采用了与 Giuliano 等[3]完全相同的文档级十倍交叉验证策略，这样可以最大化地利用实验资源，也利于实验结果与前期相关研究进行对比。我们采用的评价标准是关系

3 http://mars.cs.utu.fi/PPICorpora/GraphKernel.html

4 http://svmlight.joachims.org/

抽取中普遍采用的准确率（P），召回率（R）和 F 测度（F1），另外，AUC（area under the receiver operating characteristics curve)可以衡量数据类别在任何分布或任何错误代价下分类算法的总体性能，已经广泛用于机器学习中对分类算法进行评价。因此我们也提供了与 Airola 等[6]和 Miwa 等[9]的 AUC 性能比较。

4.2 各种特征的表现

表 1 显示了在 AIMed 语料库上，采用 10 倍交叉验证策略时，各种特征对系统性能的影响，为了显示依存树中动词特征对性能的贡献度，我们将依存特征分为两部分进行实验。

表 1 不同特征对 PPI 抽取的影响

特征	P(%)	R(%)	F1
词汇信息	59.4	40.6	47.6
+基本短语块	59.2	44.5	50.6
+句法树	60.9	44.8	51.4
+DependencySet1	62.9	48.0	53.9
+DependencySet2	**63.4**	**48.8**	**54.7**

表 1 显示本系统在准确率，召回率和 F 测度上分别达到 63.4%，48.8%和 54.7，也表明了以下特点：

- 词汇特征获得的性能比较低，尤其是召回率只有 41%左右。这表明单靠词汇特征本身不能准确表达 PPI 的关键信息，也说明 PPI 抽取的难度较大。
- 相对于词汇特征，基本短语块特征将召回率提高 3.9%，F 测度提高 3。可以看出基本短语块特征是获取局部句法信息的重要特征。
- 句法树特征的作用不很明显，它仅提高 F 测度 0.8 个单位，造成这种情况的原因之一可能是蛋白质实体之间的路径往往过长，产生了数据稀疏问题；另一方面，句法树特征 PTP 有时会部分包含在基本短语块特征 CPP 中，此时句法树特征就没有为系统提供新的可用信息。
- 依存特征 DependencySet1 也十分有效，它将准确率和召回率分别提高了 1.6%和 2.3%，F 测度也随之提高了 2.5 个单位。这表明依存特征可以有效捕获 PPI 实例，且能避免浅层句法信息中经常出现的噪音。统计数据显示 AIMed 语料库中蛋白质实体距离大于 5 个词的句子占总数的 60%以上。所以，依存特征在 PPI 抽取中具有巨大潜力，因为它们可以抽取长距离的依存信息。以上面提到的句 1 为例，虽然两个蛋白质在句子中相距较远，但它们之间的依存关系却简明而清晰，表达了相互作用的信息。
- 依存树中的动词特征提高 F 测度 0.8 个单位，这是因为一些动词如 interact、active 和 inhibit 等，能强烈暗示两个蛋白质实体的关系，为检测 PPI 提供了可靠的信息。

4.3 与其他系统的比较

表2是本系统与其他主要PPI抽取系统性能的对比，表中仅列出了采用相同实验设置的系统。按照不同的机器学习方法，我们将所有的系统分为三类：基于特征向量的方法，基于核函数的方

法和基于复合核的方法。表 2 显示了 Airola 等， Miwa 等和 Kim 等采用基于核函数的方法获得了相对高的性能，但本系统获得的 54.7 的 F 值是所有基于特征向量的方法中最好的，即使与某些基于核函数的方法相比也处于领先水平。

表 2 与其他系统的比较

系统	P(%)	R(%)	F1
基于特征向量的方法			
本系统	63.4	48.8	54.7
Giuliano 等（2006）[3] [5]	60.9	57.2	59.0
Mitsumori 等（2005）[1]	54.2	42.6	47.7
Yakushiji 等（2005）[11]	33.7	33.1	33.4
基于核函数的方法			
Kim 等 （2010）[7]	61.4	53.3	56.7
Airola 等（2008）[6]	52.9	61.8	56.4
Bunescu 等（2006 ）[12]	65.0	46.4	54.2
基于复合核的方法			
Miwa 等（2009a）[9]	-	-	62.0
Miyao 等（2008）[8] [6]	51.8	58.1	54.5
Sætre 等（2007）[5]	64.3	44.1	52.0

为了测试本系统在生物医学语料库上的泛化性能，我们也在 BioInfer、HPRD50、IEPA 和 LLL 四个 PPI 语料库上用同样的方法进行了实验。表 3 显示了相应的 F 值，AUC 测度及其标准差，并与 Airola 等[6]和 Miwa 等[9]的数据进行了对比。

表 3 在其他 PPI 语料库上的性能

	本系统				Airola 等 (2008) [7]				Miwa 等 (2009a)			
语料库	F1	σF1	AUC	σAUC	F1	σF1	AUC	σAUC	F1	σF1	AUC	σAUC
AIMed	54.7	4.5	82.4	3.5	56.4	5.0	84.8	2.3	60.8	6.6	86.8	3.3
BioInfer	59.8	3.5	80.9	3.3	61.3	5.3	81.9	6.5	68.1	3.2	85.9	4.4
HPRD50	64.9	13.4	79.8	8.5	63.4	11.4	79.7	6.3	70.9	10.3	82.2	6.3
IEPA	62.1	6.2	74.8	6.6	75.1	7.0	85.1	5.1	71.7	7.8	84.4	4.2
LLL	78.1	15.8	85.1	8.3	76.8	17.8	83.4	12.2	80.1	14.1	86.3	10.8

表 3 显示我们的系统性能与另外两个系统的趋势基本一致，LLL 语料库上均获得最好性能和最大的 F 值标准差，而在 AIMed 的性能都是 5 个语料库中最差的。

[5] Airola 在正确的数据集上重现了实验，得到的 F 值为 52.4

[6] 此处的数据是[7]中与本系统实验设置最接近时的结果（SD 表示）。

[7] BioInfer 语料库上的 F1 和 AUC 经 Miwa 等(2009b)修正。

5. 结论与展望

本文以 SVM 为分类器，用统计学习的方法实现了一个有指导的 PPI 抽取系统并且获得了 F 测度 54.7 的较好性能。本系统综合研究了多种词汇，句法尤其是依存特征对 PPI 抽取的影响。我们发现依存树特征和基本短语块特征对 PPI 抽取的贡献最大，依存树中的动词特征能进一步提高系统性能。另外，在多个生物医学领域语料库上的实验也检验了本系统的泛化性能。

下一步工作中，我们将在基于特征向量的 PPI 抽取中探索更多的句法特征，以进一步提高系统的性能，同时也将考虑将平面特征与结构化信息更好结合起来的途径。

参 考 文 献

[1] T. Mitsumori, M. Murata, Y. Fukuda, K. Doi, and H. Doi. 2006. Extracting protein-protein interaction information from biomedical text with SVM. IEICE Transactions on Information and Systems, E89-D (8): 2464–2466.

[2] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. 2003. Extracting information on protein-protein interactions from biological literature based on machine learning approaches. *Journal of Genome Informatics*. (14): 699–700

[3] C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proceedings of EACL'06*, pages 401–408. April, 2006. Trento, Italy.

[4] G.D. Zhou, J. Su, J. Zhang, and M. Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL'2005*, Ann Arbor, Michgan, USA 2005, pages 427-434.

[5] R. Sætre, K. Sagae, and J. Tsujii. 2007. Syntactic features for protein-protein interaction extraction. In *Proceedings of LBM'07*, volume 319, pages 6.1–6.14.

[6] A. Airola, S. Pyysalo, J. Bjõrne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross corpus learning. BMC Bioinformatics.

[7] S. Kim, J. Yoon, J. Yang and S. Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *Journal of BMC Bioinformatics*, 2010 11:107

[8] Y. Miyao, R. Sætre, K. Sagae, T. Matsuzaki, and J.Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL-08*: HLT, pages 46–54.

[9] M. Miwa, R. Sætre, Y. Miyao and J. Tsujii. 2009a. Protein-Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers. *Journal of Medical Informatics*, 78(2009): e39–e46

[10] M. Miwa, R. Sætre, Y. Miyao and J. Tsujii. 2009b. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. In *Proceedings of EMNLP'09*, pages 121–130. August, 2009. Singapore.

[11] A. Yakushiji, M. Yusuke, T. Ohta, Y. Tateishi, J. Tsujii. 2006. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of EMNLP'06*, Sydney, Australia 2006, pages 284–292.

[12] R. Bunescu and R. Mooney. 2005. Subsequence kernels for relation extraction. In *Proceedings of NIPS'05*, pages 171–178. December 2005.