生物医药语料库词汇分析

杨瀚轶[1], 王世松[2]

[1]华中农业大学信息学院，430070，武汉，湖北，中国
[2]华中农业大学信息学院，430070，武汉，湖北，中国

摘要

语料库是存储在计算机上，用于研究语言是如何使用的书面或口头的自然语言材料集合。更准确地说，语料库是用于语言分析和语料分析的系统化和计算机化的真实语言集合。为了开发NLP应用，我们需要书面或口头的自然语言材料作为语料库。这些材料或数据被用作输入并帮助我们开发NLP应用。语料库是NLP相关的应用中最关键和最基本的部分，它提供了用于构建应用的定量数据。语料分析可以被定义为一种以真实上下文和交际语境为基础的，深入研究语言方法。本节讨论的是数字化存储的，可以通过计算机获取、检索和分析的语料库。

关键词: NLP，语料库，TTR，词云

1 课题概况

文本数据的语料分析包括对数据集的统计调查、操作和泛化。对文本数据集，我通常对语料库进行出现多少不同的单词，单词频率分别是多少的分析。几乎每一个NLP应用开发都需要进行一些基础的语料分析来帮助我们更好地理解语料库。

近些年来，深度学习，人工神经网络在我们生活中频繁出现，NLP也发展到了新高度，我们粗略的了解NLP分析的一些流程，认为对语料库的处理和分析是NLP后续操作的基础，这对于我们未来NLP的扩展学习和同专业知识结合有较大的价值。我们也通过对生物医学语料库的分析，掌握了文本处理的一点基本方法。

2 数据

为了对比生物医学类语料库与生物类著作在用语上的差异，本次试验我们共收集了五个语料库，分别是三个生物医学类语料库和两个我们自己构建的生物书籍类语料库，生物医学类语料库包括AIMED、GENEREG、IEPA，生物书籍类语料库包括由《物种起源》、《GENEX》、《生命是什么》构成的语料库，我们定义为严谨型生物类书籍语料库，由《昆虫记》、《昆虫记忆》、《自私的基因》构成的语料库，定义为科普型生物类作品语料库。

3 研究方法

3.1 研究方法的算法背景，与其他方法的联系与区别

NLP是从1950年图灵测试才出现，图灵测试指的是通过人与机器交流让人判断交流的是人还是机器，验证机器是否智能。在1950-1970年间，NLP的主流是基于规则形式的语言理论，科学家根据数学中的公理

化研究自然语言，试图用有限的规则描述无限的语言现象。1970年至今，NLP的主流是基于统计，Google机器翻译打败了基于规则的Sys Tran [1]。2010年后，机器学习逆袭成为主流，AlphaGo掀起了人工智能潮。

3.2 研究方法中的核心思路

我们搜索资料，了解到NLP的一般流程主要包括获取预料，语料预处理，特征工程，特征选择，模型训练，评价指标以及模型上线应用，这次实验我们主要进行的是前两个步骤。通过语料清洗，分词分句处理，去停用词等计算语料库的TTR，绘制语料库的词云以及对语料库的句子进行比较 [2]。

3.3 本文的方法部分与课堂讲授内容的联系和区别/或补充

本文的基本代码与课堂讲授内容几乎没有差别，但是在运行过程中为达到不同目的有少量修改，对于整体实验的思路，我们也是在课堂讲授的基础上进行拓展并自己加以分析与推测。我们使用到的公式是TTR=$\frac{amount\quad of\quad unique\quad words}{amount\quad of\quad total\quad words}$,，用以计算词汇的多样性 [3]。

4 算法实践和代码编写要求）

4.1 任务描述

收集5个生物医学类语料库，对这5个语料库进行一系列分析，分析包括课堂所讲述的TTR与词云，以及尽可能拓展分析。

4.2 实验设计

此次实验使用了Linux，R-4.0.5，Python，在Linux中主要进行的是对数据的预处理，在R中主要进行的是检验与绘图，在Python中仅使用了自己编写计算句子词数的代码。对于实验数据的预处理，我们进行了分词处理，词性还原和分句处理。

在Linux中，通过将语料库的非字符全部转换为换行符，将语料库内所有大写字母转为小写字母，将处理后的语料库进行排序三步操作将语料库进行了分词处理并存入新文本。

因为语料库中多数单词有不同的形式，我们对分词处理后的语料库文件进行词形还原。通过安装TreeTagger包，使用tree-tagger-english命令对文件进行了词形还原并存入新文本以备用，针对TTR的比较时，我们认为词形的不同也能体现用词的丰富度，因此未作出词形还原处理。

在词汇层面处理完成后，为了解每个语料库之间的句子长度差异，我们对语料库原文件做了分句处理，原理同分词处理。通过tr命令，将语料库原文文件中的分号”;”和句点”.”作为分句标准 [4]，使它们转变为Linux系统可识别的”\ n”，存入新文本。

5 主要的生物信息学实验和实验结论

5.1 语料库间TTR的比较

在Linux中，从语料库整体的TTR入手，通过wc命令分别查看分词处理后文件的总词数，得到严谨型语料库总词数为854816，科普型语料库总词数为761228，AIMED语料库总词数为49168，GENEREG语料库总词数为77297，IEPA语料库总词数为14925，将其分别作为各自TTR计算的分母；通过sort命令-u参数的处理后，使用wc查看去重后的总词数并将其作为各自TTR计算的分子，得到严谨型语料库总词数为31849，科普型语料库总词数为32149，AIMED语料库总词数为6407，GENEREG语料库总词数为7862，IEPA语料

库总词数为2998，计算各自的比值，获得了5个语料库整体的TTR值。为了减少实验误差以及增加结果的可信度，需要再对5个语料库进行随机抽样后假设检验，后比对5个语料库之间的TTR差异。

在R中，载入dplyr包，通过sampel函数对导入的文件进行随机抽样，每个均抽取5000组，记录每次抽取到的总词数和通过distinct函数去重后的总词数便于后续计算随机抽样得到的TTR值，由于每次抽取的词数由每个语料库的总词数决定，为了使抽取到的TTR值更接近整体TTR值从而提高准确度，我们每次抽取的词数为每个语料库总词数的85%。之后对每组5000个抽样的TTR值进行了正态检验，发现只有IEPA语料库的抽样是不符合正态性的，而其它四个语料库的抽样结果都有不同程度的符合正态性，这是不合理的，我们抽取出的TTR结果应比较平稳，不会不符合正态分布，所以通过查询资料得到了结果，我们找到了一可以说明这个现象存在的例子，并且这个例子的作者还简述了这个现象说明我们不能拒绝样本来自于正态总体的原假设：“我们取有序的正整数序列[1:30]进行Shapiro正态性检验，众所周知，正整数序列完全不是正态的”，而我们看到的结果是p值为0.2662，符合正态分布，作者又说“我们看到正整数序列[1:30]不能拒绝原假设，但它绝不是正态的，可以看到，p值大于0.05的显著性水平，但是我们不能证明样本数据就是服从正态分布的，只能说不能拒绝样本来自于正态总体的原假设”，所以我们抽样的TTR结果符合正态分布可能是因为这些数据能够存在于某个正态总体。在语料库之间执行var.test()函数，得到它们之间方差不同质的结果。根据正态性检验和方差同质性检验的结果，我们调整t.test()的参数，对每组语料库之间进行了t检验，发现它们之间均存在极显著差异，绘制更直观的箱型图，展现了每个语料库抽样TTR的差异（图1 a.）。 计算每个语料库抽样TTR的平均值，得到如下结果：严谨型书籍语料库的整体TTR为0.0372583，抽样获得的平均TTR为0.040917；科普型书籍语料库的整体TTR为0.0422331，抽样获得的平均TTR为0.04614521； AIMED语料库的整体TTR为0.1303083，抽样获得的平均TTR为0.1409662；GENEREG语料库的整体TTR为0.1017116，抽样获得的平均TTR为0.1105369；IEPA语料库的整体TTR为0.2008710，抽样获得的平均TTR为0.2154706。

抽样获取的平均TTR比整体TTR都要稍大一些，但语料库之间的差异关系依旧没有改变，并且在进行语料库的TTR分析时，我们发现词数较少的三个语料库的TTR均显著大于由书籍构成的字数较多两个语料库，所以我们不禁猜测由于用词的限制，导致当词量达到一定程度后，将出现很多的重复词汇，导致了TTR的下降，因此现在看起来由书籍构成的语料库的用词丰富度更低，为了验证这种猜想，我们将六本书分开，每本书作为一个语料库进行TTR的计算。我们使用同样的抽样方法进行抽样计算TTR，通过计算得到如下结果：《GENEX》的整体TTR为0.0399536，抽样获得的平均TTR为0.04384342；《物种起源》的整体TTR为0.0526115，抽样获得的平均TTR为0.05783426；《生命是什么》的整体TTR为0.1124671，抽样获得的平均TTR为0.122679；《昆虫记》的整体TTR为0.0306737，抽样获得的平均TTR为0.03380502；《自私的基因》的整体TTR为0.0872458，抽样获得的平均TTR为0.09440558；《昆虫记忆》的整体TTR为0.1093851，抽样获得的平均TTR为0.1186744。

同之前合并的TTR数据比对发现，每本书籍的平均TTR比合并TTR略高一点，并且六本书籍分开作为语料库后TTR确实比合在一起时的TTR明显要更高，我们基本验证了我们猜想是成立的，因此我们认为：语料库的TTR也会与语料库总词量相关，当语料库总词量达到一定程度后会受限于词汇的使用而出现大量重复词汇，导致TTR下降。我们进一步思考：语料库的TTR也能从一定程度体现了语料库的好坏，在保证词汇的正常使用和多样性的同时，构建语料库不宜采用过多的词量，对语料库的后续操作效果也会较好。这也是我们自己构建语料库的一个弊端，书籍构建的语料库词量过大，但词汇的丰富度又远远不足，导致了我们构建的两个语料库TTR较低的现象。

为了更清楚直观地感受生物类书籍语料库与生物医学语料库之间的差异，我们又用同样的方法，将上述6本书以及3个语料库分别作为语料库进行了对比，绘制了一张结果较为鲜明的箱型图（图1 b.），可以得知单一生物类书籍与生物医学语料库的用词差异也是非常显著的。

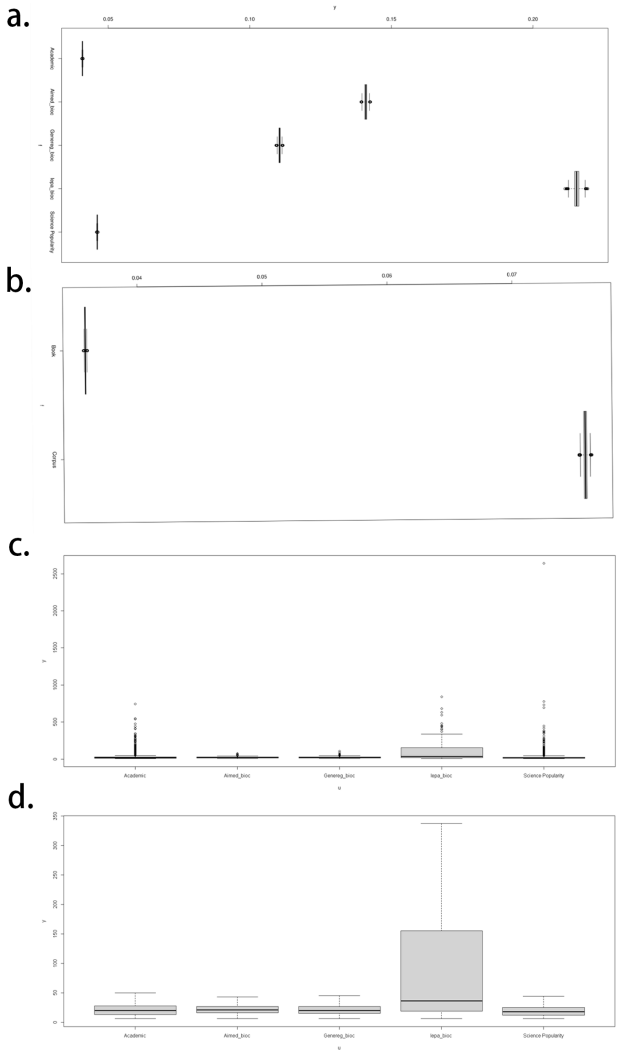

图1:
a.体现我们所使用的5个语料库的TTR差异的箱型图
b.为了进一步验证书籍和文献等的用词差异而作的我们使用的6本书籍与其它3个语料库的TTR差异的箱型图
c.我们此次所使用的5个语料库句子长度差异可视化的箱型图（未去离群值）
d.我们此次所使用的5个语料库句子长度差异可视化的箱型图（已去离群值）

图 1: 箱型图

5.2 词云的绘制

在R中导入词形还原后的语料库文件，通过课上老师讲解的代码结合课后网上的教程，我们载入tm、SnowballC、wordcloud、RColorBrewer包进行后续实验。

对读入的文件使用tm_map()函数，调整参数删除特殊符号，英语停用词，多余空格等操作后，用TermDocumentM数生成了文档矩阵后，对矩阵内的数据进行了降序排序，生成了词频数据框，最后使用wordcloud()函数进行了词云的绘制，得到了五个语料库的词云图（图2）。

只是词云并不能让我们看出每个语料库中详细的词频差异，因此我们将每个语料库的词频数据进行了可视化，得到了5个语料库的词频分布图，可以看出语料库使用次数较多的词几乎集中在前两个词，AIMED语料库词频较高的词最多，有5个词在该语料库中出现频率较显著，对于我们研究的生物医学类语料库，除去较显著的词后，剩余词的使用频率都相差不大，而书籍里出现频率较显著的词更少，只有一两个，几乎可以断定一个语料库使用较多的词与该语料库的中心有关。

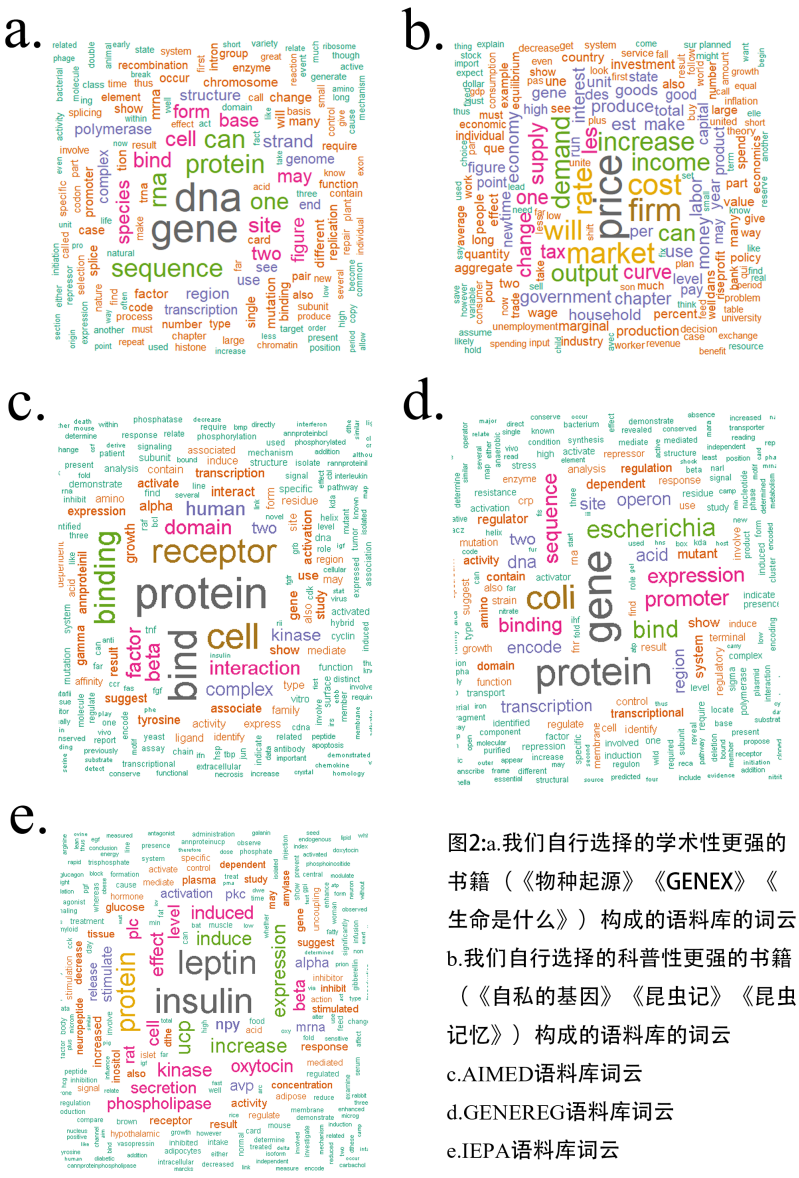

图2:a.我们自行选择的学术性更强的书籍（《物种起源》《GENEX》《生命是什么》）构成的语料库的词云
b.我们自行选择的科普性更强的书籍（《自私的基因》《昆虫记》《昆虫记忆》）构成的语料库的词云
c.AIMED语料库词云
d.GENEREG语料库词云
e.IEPA语料库词云

图 2: 词云

5.3 语料库句子长度分析

编写python代码，把分句处理后的文本导入后按行输出每句话的词数，作为5个语料库句子长度分析比较差异的基础数据，数据导入R中后，去除数据中数值小于等于5的数据，对剩下的数据进行不同语料库之间句子长度差异的分析。提取之后各语料库所剩行数如下：严谨型：34470；科普型：35133；AIMED：2011；GENEREG：3356；IEPA：113。

对每个语料库的句子长度数据进行抽样，我们查看了每个语料库的数据量，即筛选后所剩的句数，考虑到每次抽取的数据的数量相同以及数据抽少了对每个语料库的有效性不同，所以我们决定按照最大数量的语料库句数作为抽取数的15%并作小幅度提升，使用每个语料库抽取250000次，每次抽取的数量均为1的方法进行sample随机抽样，然后将数据分别存入事先准备好的空数据框进行后续分析。由于抽取的数量较多，此处耗时较长。

将抽样好的数据先进行一次正态性检验，发现5个语料库的句子长度都极不符合正态性分布，又对其进行了F测验，查看它们之间的方差同质性，检验后的p值均小于2.2e-16，证实这些语料库的句子长度两两

之间方差均不同质，所以可以进一步进行t.test()，并将参数修改为“paired=FALSE，var.equal=F”，然后来检验它们之间是否存在显著差异。结果显示，数据两两之间拥有极显著差异，由于p值小到一定程度后t.test()不会显示具体p值，无法直观了解它们之间的差异，所以我们通过作图将结果更直观地展现出来。

直接绘制箱型图会因为有大量的离群值（图1 c.），导致绘制的结果并不好，因此我们尝试对离群值进行处理。通过在boxplot()中添加参数” outline = FALSE”，在绘制箱型图时去除离群值（图1 d.）。对比两次箱型图，除了IEPA语料库的句子长度数值分布较广泛以外，其他4个语料库的结果差异并不显著，所以我们在这里初步猜测之前t.test()的结果所表明的所有语料库两两之间差异极显著很有可能是由大量的离群值所导致的。观察去除离群值前的箱型图，我们能够明显看出IEPA语料库句子长度的分布最广，离群值偏多，AIMED语料库和GENEREG语料库的离群值较少，分布较为集中，自行构建的严谨型语料库和科普型语料库有较多离群值，并且科普型语料库的句子长度离群值波动很大，这些从一定程度上支持着我们对“t.test()结果与箱型图直观感受不太一样的原因”的猜测。因此我们通过箱型图识别的方法去除了原数据中的所有离群值，然后又通过相同的方法对每组数据之间进行了一次t检验。

使用去除了异常值的数据进行t测验后，结果依旧表明5个语料库两两之间存在极显著差异，那么上述猜测就被推翻了，并不是由大量离群值造成的，于是我们猜测可能是由于IEPA语料库句子长度数值分布较广，所以在图像上展现出的结果显得其余4个语料库分布都较为集中，从而让我们觉得差异不显著，但这个猜测我们暂时没有想出办法去验证。

6 后记

6.1 课程论文构思和撰写过程

我们对于NLP是完全陌生的，在此次课程论文的书写中，我们秉承着以完成课上学到的知识为基础，尽我们所能进行拓展与开放思考。为了能够对NLP有基本的认识，实践前我们上网搜索关于NLP的发展历史，基本步骤，未来前景等，在了解到NLP的基本步骤后，我们意识到，此次所进行的实验只是NLP学习最基础的一步，未来对于NLP的深入学习，我们还有远远的路要走。

在语料库的选择上，我们最开始是想选择5个生物医学类语料库的，但仅仅选择生物医学类语料库对于结果的分析可能会很单一，且我们注意到目前BioNLP权威会议把事件抽取作为了主导任务，采用生物医学文献为数据源，以支持开发更细粒度，更具结构化的数据库为目的，引导人们提出各种生物事件抽取方法 [5]。所以我们在语料库的选择上采取了“3+2”，即3生物医药类语料库+2我们自己构建的生物类书籍语料库。通过文献阅读，我们了解了IEPA语料库与AIMED语料库在NLP类文献内使用频率较多，所以我们选择了这两个出现频率较高的语料库以及出现频率较低的GENEREG语料库 [6]。

确定了语料库后，我们就开始思考该如何分析语料库，除了课上所学的TTR与词云绘制，我们还可以做什么？在查阅一些资料后，发现可以分析的东西还是很多的。例如词性还原，词性标注，去停用词等。由于此次选择的语料库与情感分析，知识推理无关，所以我们在进行词性标注后发现无法继续分析便舍弃了。

具体操作过程中，有时会出现与预料不同的结果，我们会尝试对其进行扩展与发散思考，做出猜测并进行检验。

由于对语料库的操作及得出结果都是直接得到的，所以我们在撰写过程中就思考是否将每次操作后的结论单独提取出一个板块，但简单的尝试后我们发现有很多操作结果与结论可以穿插进行，单独提取出会导致实验流程部分语意不通且结论部分无序零碎。纠结过后，我们还是选择操作结果与结论穿插撰写。

使用TexWorks撰写论文时也出现了一点小插曲，保存为PDF格式时，屏幕上会提示latexpdf配置错误，查询一些解决方法，下载texlive解决了这个问题。

6.2 所参考主要资源

http://corpora.informatik.hu-berlin.de/部分生物医学类语料库下载网址

https://genialebooks.com/ebooks/生物类书籍下载网址

http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-
云绘制参考代码网址

6.3 代码撰写的构思和体会

在词云的绘制中，因为使用TreeTagger包进行词形还原时有些词未能识别而显示”unknown”结果，所以我们决定将所有”unknown”均替换为对应位置的原词，调整后重新读入R中，继续词云的绘制。

在语料库句子长度比较中，我们准备对数据进行t检验，发现Shapiro.test()函数只能检验数量在3到5000之间的数据，所以我们改变思路，采取了与之类似的，nortest包里的Anderson-Darling进行正态性检验，并且还把每组数据的前5000个数据进行Shapiro.test()来与之对比，发现其实两种方法在此次实验中并无差异。

6.4 生物信息学实验设计的构思和体会

生物信息学是一门交叉学课，在我们掌握生物知识的基础上，我们还需要熟练掌握代码的编写来适应飞速增加的生物学数据。每一次的代码编写都是一场实验，所以如何设计实验让代码完美是值得思考的，我们应在正式实验前充分了解实验的背景和前人做过的尝试，减少自己的弯路，要大胆进行尝试，在反复的试错中也许会有新的发现或者新的想法，新的发现和想法要尽可能验证，不断的推翻与建立，我们才会学到更多的知识，逐渐接近正确的方向。

6.5 人员分工

为了达成人均得到练习的目的，我们采取了双线程并行的方式，共同完成了实验数据预处理、词云绘制、TTR分析、句子长度分析、文献查找、论文撰写等一系列工作。其中，实验数据预处理由二人对半完成，TTR分析，词云绘制，句子长度分析由二人对半完成大部分工作，在最后处理上，杨瀚轶负责TTR与句子长度分析的整合和比对，王世松负责词云的整合和比对以及对句子长度分析的复查，文献查找由二人共同商讨搜集，论文撰写部分由杨瀚轶完成第5部分与人员分工部分的撰写，其余部分由王世松完成，最终复查由二人共同完成。

参考文献

[1] 博客园. 自然语言处理（nlp）的发展. Accessed 21 May 2021. https://www.cnblogs.com/mantch/p/11385113.html, 2014-8-20.

[2] 百度. 自然语言处理的一般流程. Accessed 21 May 2021. https://baijiahao.baidu.com/s?id=1651906999137532506, 2019-12-03.

[3] G McKee, D Malvern, and B Richards. Measuring vocabulary diversity using dedicated software. In *Literary and Linguistic Computing*, pages 323–328, 2000-9.

[4] 刘敏捷. 基于组合学习和主动学习的蛋白质关系提取. PhD thesis, 大连理工大学, 2015-6-8.

[5] 王健. 面向生物医学领域的信息提取的关键技术研究. PhD thesis, 大连理工大学, 2014-5-16.

[6] 郭瑞. 基于迁移学习和词表示的蛋白质交互关系抽取. Master's thesis, 大连理工大学, 2015-5-5.