硕 士 学 位 论 文

基于语义和半监督学习的医学文献知识发现

Knowledge Discovery from Biomedical Literature Based on Semantic Resources and Semi-supervised Learning

作 者 姓 名：李宗耀

学科、专业 ：计算机应用技术

学　　号 ：21109224

指 导 教 师：杨志豪 副教授

完 成 日 期：2014.5.3

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作者签名：_____ 日期：_____年____月____日

摘　要

目前，每年生物医学文献的数量正在呈指数的方式增长，科研人员为了得到好的研究成果，需要查找阅读大量的文献，但如此大规模的文献，给科研人员带来了巨大的困难。同时，现代科学研究分工明确，不同学科之间的交流匮乏，交叉学科的知识往往会被人们忽视，而这些文献中隐含着大量有用的、潜在的信息。Swanson在1986年提出了基于非相关文献的假设发现研究，提出并验证了鱼油可以治疗雷诺氏病的经典案例。随后许多研究人员对假设发现做了大量研究，并取得了大量的研究成果。

但传统的基于简单共现的方法会产生大量的目标词，导致很难发现有用的假设。本文提出了一种基于语义资源的方法，利用SemRep工具抽取句子内实体之间的关系，结合语义类型、概念的信息量以及关联规则对连接词、目标词进行过滤，并根据统计量信息对目标词进行排序。通过对Swanson发现的经典病例进行验证，实验结果表明该方法取得很好的效果。

另一方面，由于SemRep工具产生的语义关系准确率召回率(55%)比较低，会丢失文本中大量的关系，并影响最终的发现结果。本文使用基于词特征的核和图核来抽取句子之间的关系，并使用半监督学习Co-training的思想对训练集进行扩充，在句子关系抽取方面相较于SemRep有提高。本文使用以上的关系抽取方法，利用SVM分类器，分别建立了AB、BC两个监督模型，对于不同语义类型的关系分别进行抽取，并与SemRep工具在经典病例上进行对比，实验结果表明该方法取得较好的效果。

关键词：隐含知识发现；非监督学习；语义关系；协同训练

Knowledge Discovery from Biomedical Literature Based on Semantic Resources and Semi-supervised Learning

Abstract

Nowadays, with the rapid growth of biomedical literatures, it is difficult for biological researchers to find their related literatures from amounts of biomedical literatures. Meanwhile, there are few communications between different science disciplines as different disciplines have their own focus. It is easy to ignore the useful and potential information in the interdisciplinary field. Swanson(1986) first found a connection between Fish Oil and Raynaud's syndrome in disjoint literature areas. The purpose of Swanson's pioneering work is to find links that are not previously founded by researchers, and he defined the process as Literature Based Discovery (LBD). After him, many researchers put a lot of efforts to discover hypothesis links in disjoint literature areas.

However, using traditional methods based on co-occurrence is difficulty to find useful and valuable information, because those methods produce too many target concepts. This paper presents a method based on semantic resources, using SemRep system to extract relationships between entities within the sentence. By employing a combination of the semantic type, concept information amount and association rules, it is effective to filter the linking and target concepts and sort the target concepts with the statistical information. The experimental results demonstrate that our method works well on the classic cases found by Swanson.

On the other hand, the semantic relations extracted by SemRep are not comprehensive, as the recall rate is only 55%, it may lose a lot of useful information. This paper presents a method which is based on bag of words and kernel of graph to extract the relationship between the two concepts in a sentence. The co-training is used to expand the training set for getting a good model. Compared to SemRep, this paper's method perform well. By using the SVM classifier, we develop two models which are the model of AB and BC respectively. Compared with SemRep, the two models perform better on the classic cases found by Swanson.

Key Words： Literature Based Discovery; Semi-supervised Learning; Semantic Relation; Co-training

目　　录

1 绪论

1.1 研究背景及现状

1.1.1 研究背景

在现代科学研究中，由于各个学科之间分工明确，大部分研究人员只关注自己的学科，对于别的学科投入的精力有限，缺乏不同学科之间的交流，进而导致一些涉及交叉学科的知识被忽视。但是科学上的新理论、新发明的出现，通常是在不同学科的交叉领域。科研人员对交叉学科的重视，能够推动科学研究的发展与进步。过去的科学，只有数理化等六个一级学科，而经过多年的发展，新形成了许多交叉学科，如生物学与化学交叉形成了生物化学、计算机学与数学交叉形成了计算机科学、生物学与信息学交叉形成了生物信息学等。但是各自学科的文献数据不能被及时的发现并提取出来用到交叉学科领域中，可能在某一学科中一种科学现象已经在很多年前就得到了验证，但是由于缺乏信息交互，在另外的一个学科中却无法了解这一实事。

另一方面，由于现在科学文献的数量每年以惊人的增长速度，甚至在同一学科中也存在不同文献之间的关联知识有可能被忽视的问题。例如，在生物医学领域，生物医学文献的数据库 MEDLINE 中的文献已经超过二千万[1]，同时每年也有几百万篇的文献新加入到数据库中。如此海量的数据给研究者们带来丰富的信息，但阅读如此大量的文献对于医学研究者来说是相当困难的。因此，利用文本挖掘的方法自动地从生物医学文献中提取和组织有用的信息变得尤为迫切。

1986 年，美国芝加哥大学 Swanson 教授提出了基于非相关文献的知识发现[2]。其主要思想是从两篇不同的、非相关的文献提取知识片段，将知识片段组合到一起，发现新的知识，并能够从逻辑上对新发现的知识进行解释。其发现过程如下：考虑两个不相关的文献集 CL(C Literature)和 AL(A Literature)，通过挖掘文献的信息，发现文献集 CL 中提到大部分雷诺氏病患者（C）都存在一些特定的生理现象（B），如：血液粘稠度升高、血小板凝集度升高和血管收缩等；而在文献集 AL 中提到鱼油及其活性成分（A）能够降低血液粘稠度和血小板凝集度，并能使得血管舒张。因此，Swanson 得出鱼油可以治疗雷诺氏病的假设，并在之后的医学临床实验进行了验证。Swanson 的整个发现过程是知识发现中经典的 ABC 发现模型，在 ABC 模型中，C 作为初始词（Starting concept），B 作为连接词（Linking concept），A 作为目标词（Target concept），即 C 为需要治疗的疾病，A 为可能治疗这种疾病的物质或者药物。

在 Swanson 发现这种假设以前，从未有文章论述鱼油对雷诺病有治疗作用，甚至两类文献很少被共同引用或相互引用，这是两类非直接相关文献，Swanson 定义类似的研究为基于非相关文献的知识发现 LBD（Literature-based Discovery），也称假设发现。

1.1.2 研究现状

（1）国外研究现状

Swanson 在发现鱼油和雷诺氏病之间的关系后，又使用类似的方法发现和验证了偏头痛和镁、消炎痛和老年痴呆症等关系[3-7]。Swanson 和 Smalheiser 联合开发了 Arrowsmith 系统[8]，运用信息检索的技术将知识发现的过程自动化，实现了计算机软件、文献数据库和科研人员的交互。

许多学者重现了 Swanson 的研究工作，并发现了许多新的假设。早期的 LBD 研究大都采用信息检索技术，并作如下假定：如果概念 A 与概念 B 的共现次数越高，则概念 A 与概念 B 有关联的可能性越大。通过使用统计特征，自动化地实现 ABC 模型。Gordon 和 Lindsay(1996，1999)等人[9]对 Arrowsmith 中的基于单词的词频统计方法进行了改进，他们结合当时快速发展的信息检索技术，利用基于短语的词频统计方法，使用四个统计量确定名词短语潜在的价值，并对其进行排序和选择。最后 Gordon 和 Lindsay 对雷诺氏病和鱼油之间的关联进行了验证。Weeber(2001)等人[10-11]利用 UMLS 系统实现了自然语言与 UMLS 概念的映射，引入了语义类型过滤的方法，通过语义类型有目的地过滤连接词和目标词,并依据以上方法设计了 DAD 系统，使用该系统对 thalidomide（萨力多胺）这种药物的潜在用途进行了预测。Srinivasan[12]使用生物医学数据库 MEDLINE 中标注的医学主题词 MeSH 进行假设发现研究，避开了全文检索，极大地减小检索的时空复杂度，取得了不错的效果，但是使用 MeSH 检索，容易丢失医学文献本身中的有价值信息。Yetisgen-Yildiz 和 Pratt 开发了 LitLinker 系统[13]，进一步优化检索技术，使用 Z-Score、TFIDF、PMI 等计算方法来评价连接词和目标词，并设置阈值对概念进行过滤。LitLinker 系统的主要优势在于通过结合几种有效的统计特征对概念进行过滤和聚类，排除了同一概念存在多种表达方式的问题，有效地提高了知识发现过程的效率；同时该系统使用信息检索中的准确率、召回率和 MAP 等方法评测知识发现的结果。

上述方法主要使用的是基于统计的特征结合信息检索技术进行研究，普遍存在的问题是没有大规模的使用语义资源，无法对得出的新发现做合理的解释。Xiaohua Hu 等人[14-15] 利用关联规则的方法进行知识发现研究，主要通过结合语义资源，拓展了 Weeber 提出的语义类型过滤中语义数目，取得了很好的效果。Miyanishi 等人[16]在语义信息的基础上提出了事件相似度的概念，利用事件相似度来进行研究。D.Hristovski[17]利用自然

语言处理技术对生物文献进行处理，对每个句子抽取语义关系，并定义了关联规则，得到了可解释的发现。Trevor Cohen 等人[18]使用语义索引预测模型发现了许多有效的关联规则，发展了 D.Hristovski 的方法。Delroy Cameron 等人[19]在语义关系的基础上建立了图模型，拓展了 ABC 模型，并对雷诺氏病进行验证，从语义角度再现了 Swanson 的发现。D Hristovski 等人[20]通过引入基因微阵列数据，扩充了假设发现的语料范围，进一步整合利用生物信息学的资源。

（2） 国内研究现状

国内也有许多学者进行假设发现研究，主要是将假设发现应用到中医药研究或者情报学领域。刘耀等人[21]在通用语言知识库的基础上，结合中医药文献的特点，构建了中医药古文献语言知识库，为非相关文献知识发现在中医药方面的研究打下了基础。许建阳等人[22]对非相关文献知识发现对中医学发展的启示做了阐述。李文林等人[23]利用 Arrowsmith 系统对中药当归和痛经的相关性进行了分析。曹志杰等人[24]对生物医学领域与航天科技情报研究的差异性进行了对比分析，使用非相关文献知识发现的方法模拟了新型飞行器隐身技术的发现过程，对非相关文献知识发现的研究领域做了扩展。

目前，机器学习和文本挖掘的方法在生物信息学领域的应用仍处于初级阶段，需要大量科研人员的共同努力才能取得长足发展。我国生物医学领域的信息化程度不高，而且中西药之间存在差异，使得国外的一些研究方法无法直接使用。我们需要，一方面，通过建立电子病例、实时更新生物医学信息系统等手段，实现我国生物医学领域的信息化，为以后研究打下基础；另一方面，通过引入国外的研究方法，加快我国生物医学方面的研究。非相关文献的知识发现作为这一领域的经典案例，我们可以从中例汲取一些经验，以促进生物信息学的发展。

1.2 本文主要工作及章节安排

本文的研究内容重要包括两方面：基于语义资源的生物医学文献知识发现研究和基于半监督学习的生物医学文献知识发现研究。

第一章 对基于生物医学文献的知识发现的研究背景进行介绍，综述该研究中出现的方法和成果。

第二章 对基于生物医学文献的相关资源和技术进行介绍，包括生物医学文献数据库 MEDLINE、一体化医学语言系统 UMLS、生物医学概念识别工具 MetaMap、语义关系抽取工具 Semrep、医学主题词 MeSH、开放式闭合式发现算法和主要的假设发现系统。

第三章 介绍了基于语义资源的方法在生物医学文献知识发现研究上的应用，并在 Swanson 发现的经典病例上进行验证，以证明方法的有效性。

第四章 介绍了基于半监督学习方法在生物医学文献知识发现研究的应用，论述了方法的可行性，并通过实验进行了验证。

2 相关资源与技术

2.1 生物医学文献及资源

2.1.1 MEDLINE

MEDLINE[1]是由美国国立医学图书馆(The National Library of Medicine, 简称NLM)收集的国际性综合生物医学信息书目数据库，是当前世界上最权威的生物医学文献数据库，同时也是自动化的生物医学文献检索系统。该数据库涵盖了美国医学索引（Index Medicus, IM）、牙科文献索引（Index to Dental Literature）、国际护理索引（International Nursing Index），涉及的学科领域包括基础医学、临床医学、环境医学、营养卫生、职业病学、卫生管理、医疗保健、微生物、药学、社会医学等。

从 1950 年到现在，MEDLINE 数据库从 5639 种包含生物医学文献的出版物中收录了超过 2160 万条记录。这些数据可以通过互联网从 PubMed 界面免费获取，同时每周二到周六都会有新的数据被添加到数据库中。MEDLINE 数据库中的记录的主要内容是文章标题和摘要，其中 88%的文献是英文文献，76%的文献有英文文摘。

MEDLINE 可通过主题词，副主题词，关键词，篇名，作者，刊文，ISSN，文献出版，出版年，出版国等进行检索。PMro 表示文献在 MEDLINE 数据库中的唯一标识符、TI 表示文献的标题、AB 表示文献的摘要、MH 是医学主题词 MeSH。例如，PMID 为 23990406 的文献在 MEDLINE 数据库的形式如图 2.1 所示。

PMID- 23990406
TI - Migraine and the social selection vs causation hypotheses: a question larger than either/or?
AB - For decades, the question of social selection vs social causation has been raised by public health researchers and social scientists to explain the association between socioeconomic factors and mood disorders...
MH - *Causality
MH - Humans
MH - Migraine Disorders/*psychology
MH - *Social Class
MH - Socioeconomic Factors

图 2.1 MEDLINE 结构示意图

Fig. 2.1 A structure of MEDLINE article

2.1.2 UMLS

一体化医学语言系统(UMLS，Unified Medical Language System)[25]是从 1986 年开始由美国国立医学图书馆研究开发的。在生物医学文献中，存在词汇的同义多义的现象、词汇的模糊性和不确定性、词汇量巨大、词间关系不明晰等问题，使得单纯使用单一概念进行检索变得困难，无法保证检索的召回率和准确率。系统得到较好的检索效果需要检索时使用的查询语言和文献标引的词汇一致。这需要人工对文献进行标引，但是人工标引存在如下的问题：

（1）工作量大：国内外每年新增加几千万份生物医学文献，每天都会有大量的临床病例记录产生，需要标注的工作量巨大，在国外的信息检索系统中大部分的运行费用用于人工标引。

（2）一致性差：试验表明，不同人为同一主题标注不同的叙词表，会对同一文献产生叙词表不一致的问题。

（3）效率低：标注人员需要花费一个半小时以上的时间才能正确的标引一篇文献。

（4）词表的完备性低：由于文献标注工作量大和效率低等问题，使得词表的更新速度远远低于文献发表的速度，导致词表的完备性较低。

UMLS 包括四个部分：超级叙词表(Metathesaurus)、语义网络(Semantic Network)、情报源图谱(Information Sources Map)和专家词典(SPECIALIST Lexicon)，具体的结构如图 2.2 所示，本文使用的是其中的超级叙词表和语义网络。

超级叙词表主要包括生物医学概念、术语、词汇及其等级范畴。到目前为止，一共收录了 100 多万个生物医学概念，500 多万个词。这些概念和词是从 100 多个生物医学受控词表（controlled vocabularies）、分类系统（classification systems）得到的。这些生物医学受控词表主要包括 ICD-10、 MeSH、 SNOMED CT、 DSM-IV、 LOINC 等。通过涵义（meaning 或概念（concept）组织叙词表，其根本目的是将同一概念的多个不同名称和不同形式联系在一起，并给统一起来，进而识别不同概念之间的联系。超级叙词表采用三级结构模式，即概念（I 级）-术语（II 级）-词串（III 级），将同一个概念的多个不同名称和不同形式有序地组织在一起。

语义网络是为了阐述不同概念之间关系而建立的。在 UMLS 系统中，每个概念都有自己的唯一的 ID，至少有一种语义类型；不同概念之间可能存在不同类型的语义关系。在整个语义网络中，一共定义了 135 种语义类型和 54 种语义关系。语义类型形成了一个等级结构，每一语义类型有一个树状等级号，同时还被赋予一个语义类型代码，如〈T001〉、〈T101〉等。在语义网络中，分为层次结构和非层次结构两种语义关系。层次结构主要为“ISA”关系；非层次结构主要分为五大类：物理相关(physically-related-to)、

空间相关(spatially-related-to)、功能相关(functionally-related-to)、时间相关(temporally-related-to)和概念相关(conceptually-related-to)。

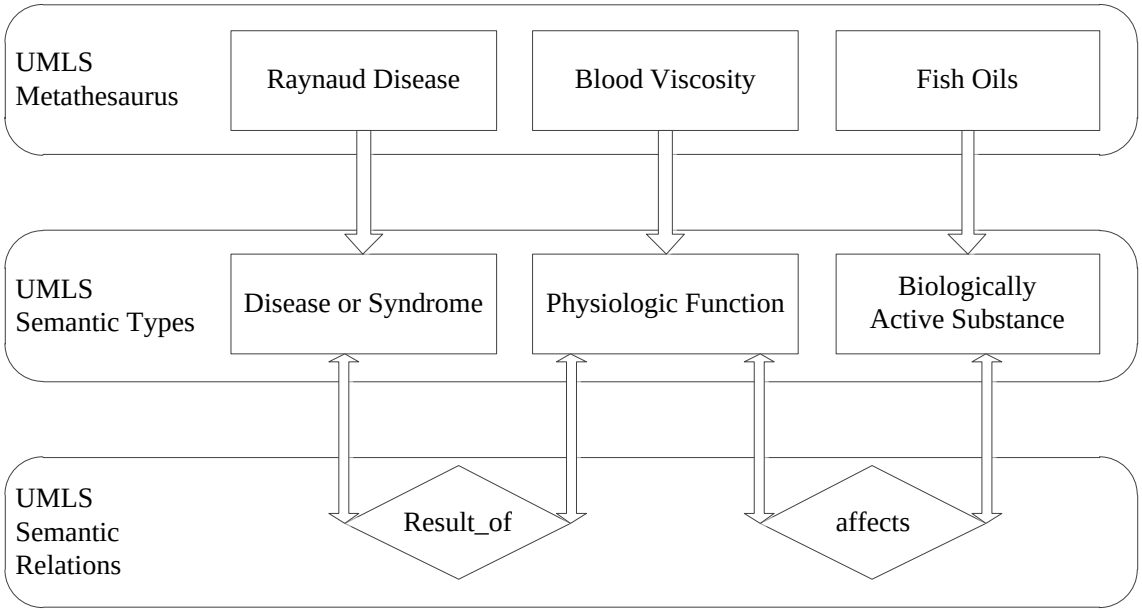

图 2.2　UMLS 系统结构图

Fig. 2.2　A structure of UMLS

2.1.3 MeSH

医学主题词（Medical Subject Headings，MeSH）是美国国家医学图书馆开发和维护的权威性主题词表[26]，是一部规范化的可扩充的动态性叙词表。MeSH 作为生物医学文献的关键词，为建立计算机文献联机检索系统提供依据，其在文献检索中的作用主要表现在两方面：专指性和准确性。MEDLINE/PubMed 生物医学文献数据库使用 MeSH 对文献进行检索。所有的 MeSH 可以从 NLM 网站中免费在线下载。

MeSH 的组成部分主要有主题词变更表、字顺表、树状结构表和副主题词表，其中主要的组成部分为字顺表和树状结构表，以下对两者进行简要介绍。

（1）字顺表　字顺表（Alphabetic List）是医学主题词表的主要组成部分。它主要由按英文字顺排列的主题词、款目词和副主题词混合组成。

（2）树状结构表　NLM 通过引入范畴表(Categories and Subcategories)使得 MeSH 具有系统性，范畴表又称树形结构(Tree Structure)，是将主题词(主要叙词)、次要叙词按照其学科性质、词义范围的上下类属关系及派生关系划为 16 个大类，分别使用 A-N、Z 表示。在 16 个大类中，有 9 类又分为若干个子类目，在子类目下面又分为若干更

小的类目。按照医学概念的性质，所有的 MeSH 都分别列入各自所属的类目之下。这样，MeSH 树状结构表能够表示 MeSH 上位词与下位词之间的关系。例如，Hepatitis B（乙型肝炎）在 MeSH 树状结构表中的关系图，如表 2.1 所示：

表 2.1　MeSH 树形结构示意图

Tab. 2.1　A tree structure of MeSH

MeSH 主题词	树形结构
Digestive System Disease（消化系统疾病）	C6
Liver Disease（肝病）	C6.552
Hepatitis（肝炎）	C6.552.380
Hepatitis,Alcoholic（乙醇性肝炎）	C6.552.380.290
Hepatitis,Animal（动物性肝炎）	C6.552.380.315
Hepatitis,Toxic（中毒性肝炎）	C6.552.380.615
Hepatitis,Viral,Human（人类病毒性肝炎）	C6.552.380.705
Hepatitis A（甲型肝炎）	C6.552.380.705.422
Hepatitis B（乙型肝炎）	C6.552.380.705.437

2.2　生物医学文献处理工具

2.2.1 MetaMap

MetaMap 是美国国立图书馆（NLM）建立的一个自然语言处理工具，能够自动地将自由文本映射到 UMLS 超级叙词表，MetaMap 是 UMLS 系统的基石，NLM 用其对生物医学文献进行建立索引。MetaMap 文本处理的基本原理如下[27]：

（1）产生名词短语的变形词[28]：主要先对文本进行解析，将句子变成名词短语，并产生这些名词短语的变形词，主要包括名词短语的拼写变化、一次多意、一意多词和派生词缀变化等。

（2）入选叙词[29]：名词短语的所有变形词入选词串集。

（3）对入选叙词赋值[30]：计算名词短语到每一个入选的词串的映射强度，同时按照映射的强度对入选词串进行排序。

（4）构造映射：根据计算出来的结果，选择排序结果靠前的词串为对原始名词短语的 Meta 映射。

MetaMap 建立了三种不同的数据模式，用于不同程度的数据过滤。这样既可以进行高密度的语义处理，也能够用于简单浏览。三种模式如下：

（1）严格模式：适合对精确度要求比较高的任务，该模式包含超级叙词表中 53%的字串。

（2）中度模式：适合将输入文本作为一个整体而不是简单短语组成的任务，该模式包含超级叙词表中 73%的字串。

（3）宽松模式：可以获得全部的超级叙词表中的字串，适用于浏览。

如图 2.3 所示，利用 MetaMap 在严格模式下处理文本短语“raynaud disease”，会得到 6 个候选词，系统会选择得分值最高的“Raynaud Disease”作为映射结果。

```
Processing 97479605.ti.1: raynaud disease

Phrase: "raynaud disease"
Meta Candidates (7):
   1000 Raynaud Disease [Disease or Syndrome]
    861 Disease [Disease or Syndrome]
    805 MAL (MAL gene) [Gene or Genome]
    805 MAL (MKL1 gene) [Gene or Genome]
    805 MAL (TIRAP gene) [Gene or Genome]
    805 MALS (MALARIA, MILD, SUSCEPTIBILITY TO) [Finding]
Meta Mapping (1000):
   1000 Raynaud Disease [Disease or Syndrome]
```

图 2.3 MetaMap 映射示例

Fig. 2.3 An example of MetaMap

2.2.2 SemRep

SemRep 是一个能够自动从文献中识别生物医学概念以及概念之间关系的自然语言处理系统，是一个基于 UMLS 的生物医学文本处理工具。通过 MetaMap 将文本中的名词短语映射成概念，依据专家词典（SPECIALIST Lexicon）和 Xerox 词性标注器[31]进行语义关系分析。具体过程如下：

（1）将句子分块，形成名词短语，利用 MetaMap 将这些名词短语映射成超级叙词表中的概念。

（2）分析名词短语的语义类型，以及句子中的词语的词性结构。

（3）将句子中的可靠关系进行输出，形成预测关系。

下面以句子“We conclude that diltiazem is effective in the treatment of Raynaud's phenomenon, especially in patients with idiopathic vasospastic disease.”为例，阐述 SemRep

的工作过程，如图 2.4 所示，在过程（1）中，对句子进行分割；在过程（2）中，对名词短语进行命名实体识别；在过程（3）中得到概念之间的关系。

We conclude that diltiazem is effective in the treatment of Raynaud's phenomenon, especially in patients with idiopathic vasospastic disease.

(1) [[We] [conclude] [that diltiazem] [is] [effective] [in the treatment] [of Raynaud's phenomenon], [especially] [in patients] [with idiopathic vasospastic disease.]]

(2) ditiazem → Diltiazem [Organic Chemical,Pharmacologic Substance]
treatment → Treatment [Functional Concept]
Raynaud's phenomenon → Raynaud Phenomenon [Disease or Syndrome]
patients → Patients [Patient or Disabled Group]
idiopathic vasospastic disease → Idiopathic disease [Disease or Syndrome]

(3) C0012373|Diltiazem|orch|TREATS|C0034735|Raynaud Phenomenon|dsyn
C0277553|Idiopathic disease|dsyn|PROCESS_OF|C0030705|Patients|podg

图 2.4　SemRep 处理结果示例

Fig. 2.4　An example of SemRep

NLM 使用 SemRep 工具对生物医学文献数据库 MEDLINE 进行处理，形成了规模庞大的语义关系数据库 Semantic MEDLINE Database[32]，用户安装 MySQL 后，可以方便地对 MEDLINE 的语义关系数据库进行检索。

2.3　算法与系统

2.3.1 知识发现算法

早期的 Swanson 对于雷诺氏病和鱼油、老年痴呆症和镁等病理的知识发现研究，都是从一个疾病开始，然后通过连接词找到治疗这种疾病的物质。这种的知识发现主要是开放式发现。Weeber 等人[10-11]对 Swanson 的研究工作进行了总结，将知识发现定义为开放式发现和闭合式发现两种方式。

开放式算法(Open discovery)：是一个产生假设的过程。在一种疾病 C 的治疗方法未知的时候，找到可能治愈或改善 C 的方法，发现过程如图 2.5 所示。

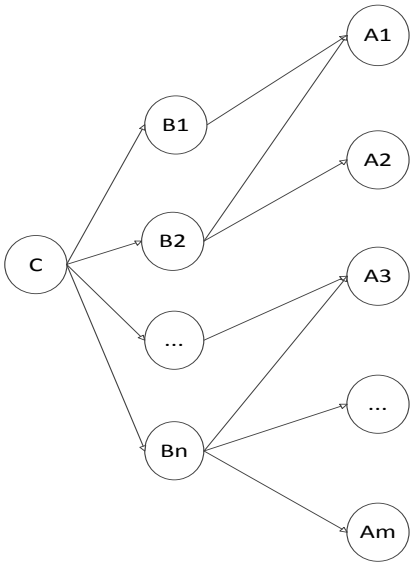

图 2.5　开放式发现

Fig. 2.5　Open discovery

闭合式算法(Closed discovery)：已知两个概念 C 和 A，当研究者去寻找两者之间的关系、阐明 A 对 C 有效的病理时，可以采用闭合式发现，发现过程如图 2.6 所示。

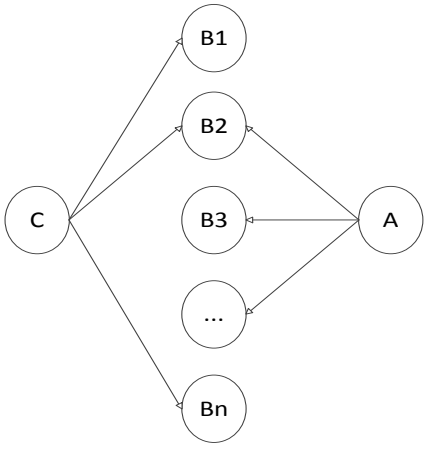

图 2.6　闭合式发现

Fig. 2.6　Closed discovery

2.3.2 知识发现系统

（1）Arrowsmith 系统

Arrowsmith 系统[8]的发现过程如下：首先，对初始词 C，检索生物医学数据库，通过对检索结果进行处理，寻找与 C 共现的连接词 B；然后，通过 B 进行数据库检索，寻找可能的目标词 A。

在整个发现过程会发现大量的连接词 B、目标词 A，系统需要对 A、B 进行过滤。Arrowsmith 系统主要的过滤手段有如下几种：

① 利用停用词列表来过滤 A、B；

② 过滤掉来源文献标题中出现的相对频率低于在整个 MEDLINE 数据库的连接词。

③ 用户可以检查连接词列表，手动剔除不合适的名词短语。

但是 Arrowsmith 只是扩展了 MEDLINE 的检索功能，并不能完全代替常规的检索，Arrowsmith 系统需要常规检索的结果作为输入的数据，进而使研究者发现文献之间的新关联和形成新科学假设。

（2）Gordon 和 Lindsay 开发的系统

Gordon 和 Lindsay(1996，1999) [9]对 Swanson 的基于单词的词频统计方法进行了改进，他们结合当时快速发展的信息检索技术，利用基于短语的词频统计方法，从获取的自然语言文本中提取 1 个、2 个、3 个词汇组成的名词短语作为文本分析对象，分析了名词短语的四个统计量：标识频次（TF）（名词短语在 MEDLINE 记录中出现的频次）、文献频次（DF）（包含该名词短语的记录数）、相对频率（RF）（名词短语来源文献中出现的频次与 MEDLINE 中的总记录数的比值）、TF*IDF（IDF 为 MEDLINE 的所有记录与 MEDLINE 中使用该名词短语的记录数的比值，并对比值去对数）。利用这四个统计量确定名词短语潜在的价值，并对其进行排序和选择。Gordon 和 Lindsay 更多的是考虑概念本身，揭示概念之间统计关系，在处理名词短语的过程中，考虑了词根变化，但是同词根的名词短语无法进行聚类，需要进行人工处理。最后 Gordon 和 Lindsay 对雷诺氏病和鱼油之间的关联进行了验证；同时通过分析四个统计量，发现相对频率的更有利于发现新的知识。

（3）Marc Weeber 等开发的 DAD 系统

Marc Weeber 等研发了 DAD 系统[10-11]，主要分析 PubMed 记录，如图 2.8 所示。该系统是一个自然语言处理系统，对假设发现的方法进行了总结，将发现过程分为开放式发现（Open discovery）和闭合式发现（Closed discovery）。开放式发现是通过初始词，利用连接词，查找目标词的过程；闭合式发现是验证初始词和目标词之间关系的过程。DAD 系统使用 UMLS 中超级叙词表中的主题概念代替文本中的名词短语。Weeber 等对系统作如下假定：

① 科研人员感兴趣的是有生物医学意义的概念，可以直接分析医学主题概念。

② UMLS 可以将文本中的名词短语映射成主题概念，通过主题概念的语义类型进行语义类型过滤。

DAD 系统的主要分析过程：首先，检索 PubMed 数据库，获取原始文本；其次，使用 MetaMap 进行主题词映射，并获得主题词的语义类型；最后，利用开放式发现或闭合式发现过程，进行假设发现研究。利用 DAD 系统，Weeber 等人也对雷诺氏病和鱼油、偏头痛和镁的关系，同时还将 DAD 系统用于发现药物副作用的研究。

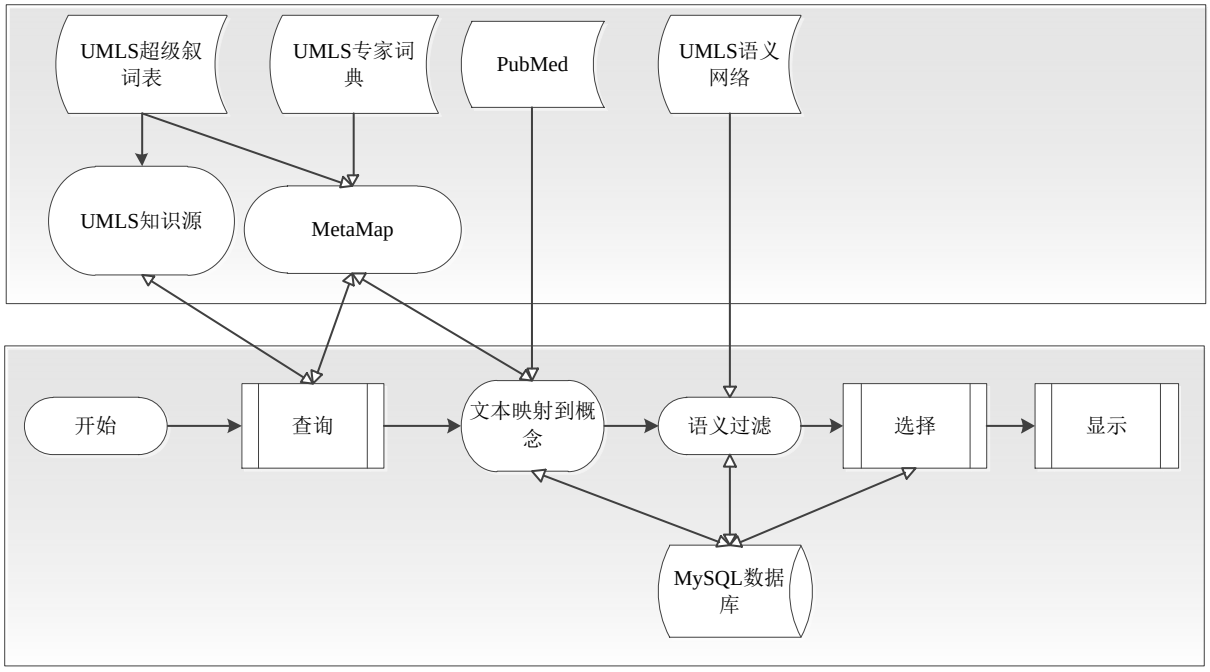

图 2.7 DAD 系统示意图

Fig. 2.7 The abridged general view of DAD system

（4）Stegmann 和 Grohmann 与 Mlink

柏林自由大学医学图书馆 Stegmann 和 Grohmann 开发了 Mlink 系统[33]，该系统的主要创新在于使用了共词分析方法和战略坐标思想，通过主题词出现的次数决定是否过滤这个主题词。

该系统的主要研究方法如下：首先利用初始词对数据库进行检索，对检索到的文献提取主题词并进行聚类分析，通过计算密度和向心度进行战略坐标图的绘制。该研究的主要的参数为：

① CDR，聚类的向心度和密度的比值。

② SIR，初始词所在文献聚类 CDR 与连接词所在文献聚类 CDR 的比值。

③ STR，初始词所在文献聚类 CDR 与目标词所在文献聚类 CDR 的比值。

初始词所在文献绘制的战略坐标图中，能够代表文献的词汇通常是 CDR 比值高的区域。Stegmann 和 Grohmann 发现那些比较重要的连接词的 SIR 值大约为 1，而比较重要的目标词的 STR 值大约为 1。该系统也验证了 Swanson 的发现，同时进行了一些新的发现研究，朊病毒和神经退行性疾病之间存在关系。

3 基于语义资源的生物医学文献知识发现研究

传统的基于简单共现的方法会产生大量的连接词、目标词，导致很难发现有用的假设；而基于语义关系的方法大都是进行闭合式发现，没有提出一个有效的对目标词排序的算法。本文提出了一种基于语义资源的方法，利用 SemRep 工具抽取句子内实体之间的关系，结合语义类型、概念的信息量以及关联规则对连接词、目标词进行过滤，并根据统计量信息对目标词进行排序。综合考虑假设发现中几种有效的方法，最大限度的对连接词、目标词进行处理，能够使有效的目标词尽可能排名靠前；同时相较于以前简单 MeSH 共现的方法，通过使用 SemRep 语料，根据实体之间的关系能够生成可解释的假设。通过对 Swanson 发现的经典病例进行验证，实验结果表明该方法取得很好的效果。

3.1 实验方法

3.1.1 语义类型过滤

表 3.1　实验所选取的语义类型

Tab. 3.1　The semantic types selected by our experiment

连接词语义类型	目标词语义类型
Biologic function	Organic Chemical
Cell function	Lipid
Finding	Pharmacologic Substance
Molecular function	Vitamin
Organism function	Element, Ion, or Isotope
Organ or tissue function	
Pathologic function	
Phenomenon or process	
Physiologic function	

在假设发现的研究方法中，语义类型过滤作为一种简单有效的连接词、目标词过滤方法，被广大研究人员所使用。Weeber[10-11]在使用 UMLS 进行假设发现研究后，就率先使用了语义类型过滤。UMLS 对其中的主题词定义了 135 种语义类型，每个 MeSH 词都有固定的语义类型，通过选择合适的连接词、目标词的语义类型，可以对二者进行有效地过滤。

在实验过程中，由于是通过计算机自动化的模拟 Swanson 进行假设发现的过程，所以初始词一般为疾病；连接词主要选取一些能够表示生理状况的语义类型；目标词作为要找的能够治疗疾病的物质，其语义类型一般为化学物质。这样就能够得到一个发现模型：一种化学物质能够改变人的生理现象，而这种生理现象又与一种疾病息息相关，因

此，这种化学物质就有可能影响这种疾病。本文主要参考 Weeber 实验中选择的连接词、目标词的语义类型，如表 3.1 所示。

3.1.2 发现模板

在本文的第二章部分介绍了句子关系抽取工具SemRep，SemRep作为一种基于规则的关系抽取工具，一共定义了54种句子中概念之间的关系。每个关系代表的含义都不一样，例如，如果概念A和B之间的关系为STIMULATES，则表示A能够刺激B；而如果两者的关系为ISA，则表示A是B的一种组成部分或者别名。这些不同的关系所包含的信息量是不同的，对于本实验，在实验中剔除了一些包含信息量比较少的关系，如“ISA”、“PROCESS_OF”、“ADMINISTERED_TO”、“ADMINISTERED_TO (SPEC)”、“PART_OF”等，保留了那些能够表明两个概念之间存在相互影响的关系，如“INTERACTS_WITH”、“INHIBITS”、“STIMULATES”等。

表 3.2　实验所选取的发现模板

Tab. 3.2　The discovery patterns selected by our experiment

C 与 B 之间的关系	B 与 A 之间的关系
ASSOCIATED_WITH	INTERACTS_WITH INHIBITS COEXISTS_WITH STIMULATES
INHIBITS	ASSOCIATED_WITH
CAUSES	COEXISTS_WITH
PREDISPOSES	INTERACTS_WITH
AFFECTS	AFFECTS DISRUPTS
COEXISTS_WITH	INHIBITS CAUSES
PREVENTS	INTERACTS_WITH INHIBITS COEXISTS_WITH

基于MeSH词简单共现的方法并没有考虑概念之间的语义关系，产生了大量无关的连接词和目标词，D. Hristovski等人[17]在SemRep等自然语言处理工具的基础之上提出了基于发现模板（Discovery Patterns）的研究方法，例如：一种药物A与一种生理现象B的关系是“INHIBITS”，即药物A能够抑制生理现象B，而生理现象B又与疾病C的关系是“ASSOCIATED_WITH”，即生理现象B与疾病C有关联,因此假定药物A能够通过抑制生理现象B来达到影响疾病的目的，也即药物A与疾病C存在关系的可能性比较大。Trevor Cohen等人[16] 使用语义索引预测模型发现了许多有效的发现模板，但是这些模板数量的

数量有限。本文在进行实验的过程中，发现会将许多有用的目标词都给过滤掉，使得目标词结果不完整。通过在Trevor Cohen等人[18]选择模板的基础上，本文对以前病例进行分析，新加入了一些模板，得到了如表3.2所示的模板集合。

3.1.3 MeSH 宽泛含义概念过滤

在整个 MeSH 主题词表中，有一些主题词，诸如：身体区域（Body Regions;A01）、细菌感染和真菌病（Bacterial Infections and Mycoses;C01）等，属于比较宽泛含义的主题词，这些主题词本身并不能提供太多的信息，但是在 SemRep 数据库中占有很大的规模，对于后期的目标词排序造成很大的困扰，因此需要对这些 MeSH 宽泛含义概念进行过滤。

在本文第二章 MeSH 部分，介绍了 MeSH 的树状结构表。MeSH 树状结构表能够表示 MeSH 上位词与下位词之间的关系。例如，MeSH“羊水”(Amniotic fluid，编号：A16.254.72)是 MeSH“胚胎”(Embryo，编号：A16.254)的一个下位词。Seco 等人[34]认为在本体中，一个概念包含的下位词越多，那么其包含的信息量越少。所以，一个概念 C 表达的信息量由下位词数量函数来表示，其计算公式如 3.1 所示：

$$IC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(N_s)} \tag{3.1}$$

其中，*hypc(c)*为概念 *c* 的下位词数量，N_s为概念 *c* 所处大类中概念的数量。

例如，MeSH“胚胎结构”（Embryonic Structures）的 N_s为其所包含下位词的数量。在公式 3.1 中，如果概念 *c* 处在 MeSH 树状结构表的顶层，则其做包含的下位词数量 *hypc(c)*必然很高，在与概念 *c* 所处大类概念的数量做对数比后，得到的值会比较大，在使用常数 1 减去比值后得到的 MeSH 主题词信息量 *IC(c)*会比较小；反之，如果一个概念的下位词数量很少，则 *IC(c)*值会很大。在实验中，对每个 MeSH 主题词计算其所包含的信息量，然后进行分析，实验保留了信息量大于 0.6 的 MeSH 词，阈值太小会导致过滤效果不明显，阈值太大会使得一些重要概念被过滤掉。

3.1.4 目标词排序

在通过开放式发现算法对文献进行知识发现后，会得到一个关于目标词的集合，但目标词的数量在经过两次共现之后变得非常大，如何对目标词排序是开放式发现的难点。Pratt 和 Yetisgen-Yildiz 等人[13]通过使用连接词的数量（Linking Term Count，LTC）对目标词排序。他们基于这样的假定：如果一个目标词可以通过多个连接词得到，就认为这个目标词与初始词的关系越密切。在他们基础之上，本文假定同一个 SemRep 关系

出现的次数越多，这个关系越可信；同时，如果一个目标词能够通过多个不同的连接词连接到初始词，也即初始词和目标词之间有多条通路，我们也认为这两者可能存在这关系的可能性比较大。在 SemRep 数据库中，相同的实体和相同的关系有可能在不同的文献中多次出现，我们利用出现的频次来评价两个实体存在关系的可靠程度，由于一个关系在数据库中出现多次后，已经是比较可靠的关系，如果直接使用频次进行计算，反而使得结果表现不佳，本文使用对数函数对频次进行处理。同时有的连接词在好多篇文献中出现，属于比较泛化的概念，提供的信息相对较少，本文利用信息检索中常用特征 DF 对其进行平滑，减小其影响因子。综合以上各个方面的因素，我们提出的量化目标词公式，如公式 3.2 所示：

$$S(A) = \sum_{i=0}^{n} \log(f(B_i, C) + 1) \log(f(B_i, A) + 1) \log(\frac{N}{df(B_i)} + 1) \qquad (3.2)$$

其中，$S(A)$为概念 A 的得分，$f(B_i, C)$为概念 C 与概念 B_i 在 SemRep 中的关系频次，$f(B_i, A)$为概念 B_i 与概念 A 在 SemRep 中的关系频次，$df(B_i)$为概念 B_i 在 MEDLINE 出现的文档频次，N 为 MEDLINE 文档集中包含的文档个数。

3.2 实验与评价

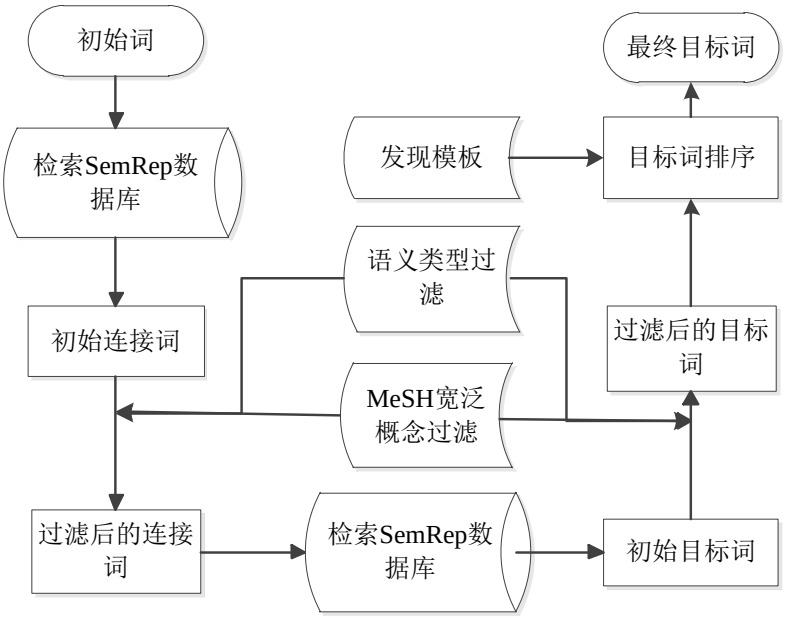

图 3.1　本实验的开放式发现流程图

Fig. 3.1　Flow chart of our open discovery approach

实验使用的语料为 Semantic MEDLINE Database[32]，这是一个已处理好的 SemRep 数据库。实验过程如下，将要查询的疾病设为初始词，利用开放式算法检索数据库，得到连接词集合，对连接词进行语义类型以及 MeSH 宽泛含义概念过滤；利用过滤后的连接词检索数据库，得到目标词集合，对目标词进行语义类型以及 MeSH 宽泛含义概念过滤，并剔除与初始词共现过的概念；然后通过发现模板寻找可能的关联，最后利用目标词排序算法对目标词排序，具体的实验过程如图 3.1 所示。

由于目前没有统一的评测标准，重现 Swanson 发现的三种经典病例（雷诺氏病与鱼油、老年痴呆症与镁、偏头痛与消炎痛）成为了生物医学领域文本知识发现评测的标准答案，以后的学者都重复 Swanson 的发现来验证他们方法的有效性，以下分别介绍三种疾病的发现结果及结果分析。

3.2.1 雷诺氏病与鱼油

表 3.3　三种疾病目标词的排名

Tab. 3.3　The rank of three diseases' target word

实验病例	目标词的排名	目标词在本语义类型下的排名
雷诺氏病与鱼油	38	3
偏头痛与镁	10	1
老年痴呆症与消炎痛	25	20

由于 Swanson 在 1986 年发现了两者的关系，为了能够重现他的实验，本文使用 1986 年以前的 SemRep 语料进行实验。为了验证本文提出方法的有效性，我们分步使用不同的过滤方法，来观察对目标词排名的影响。在实验过程中，将初始词设置为 Raynaud Disease 和 Raynaud Phenomenon（雷诺氏病和雷诺氏现象），在分步使用语义类型过滤、MeSH 宽泛含义概念过滤、发现模板过滤后，鱼油的排名依次为 274、262、38，可以看到通过模板过滤后，鱼油的排名有大幅提升,主要原因是过滤掉了许多不符合发现模板的概念；MeSH 宽泛含义概念过滤没有太明显的效果的原因是 SemRep 中的概念主要是超级叙词表，MeSH 主题词在其中仅占很小的部分，但是对于宽泛含义概念的过滤是必要的，因为在使用宽泛含义概念作为连接词进行进一步假设发现实验的时候，与宽泛含义的概念共现的目标词数量是巨大的，能够在很大程度上干扰目标词的排序，影响排序结果。鱼油在表 1 所示的目标词语义类型的排名结果为 38，而在油脂语义类型（Lipid）下的排名结果为 3，如表 3.3 所示。

我们依据找到的关系可以从数据库中抽得关系所在的句子，通过这些句子可以进行药理分析，同时十二碳五烯酸（Eicosapentaenoic Acid，EA）作为鱼油的主要成分，我

们也考察了其与雷诺氏病的关系，具体的病理分析如表 3.4 所示。我们可以看到在 PMID 为 458185 的文献中提到在雷诺氏病中血液粘稠度（Blood Viscosity）增加，而在 PMID 为 14851220 的文献中提到鱼油对血液粘稠度有影响，所以我们可以得到鱼油能够通过影响血液粘稠度进而对雷诺氏病产生影响的结论；在 PMID 为 1069430 的文献中提到雷诺氏病会伴随动脉障碍（Disorder of artery），而在 PMID 为 6320945 的文献中提到 EA 有助于改善动脉障碍，所以我们可以得到鱼油能够通过改善动脉障碍进而影响雷诺氏病的结论。具体的其他的病理分析，我们不再一一列举，从以上两个病理分析，我们能够较好的重现 Swanson 关于雷诺氏病和鱼油的假设发现。

表 3.4　雷诺氏病与鱼油的病理分析

Tab. 3.4　The pathological analysis of Raynaud Disease and Fish Oils

连接词（B）	C→B 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子	B→A 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子
Blood Viscosity（血液粘稠度）	**SemRep**： Blood Viscosity | AFFECTS | Raynaud Phenomenon **PMID**：458185 **Sentence**：A positive feedback loop between cooling, increase in digital artery resistance, increase in **blood viscosity** and a passive vascular bed is proposed as an important factor in **Raynaud's phenomenon**.	**SemRep**： Fish Oils | AFFECTS | Blood Viscosity **PMID**：14851220 **Sentence**：Beneficial effect of **fish oil** on **blood viscosity** in peripheral vascular disease.
Disorder of artery（动脉障碍）	**SemRep**： Raynaud Phenomenon | COEXISTS_WITH | Disorder of artery **PMID**：1069430 **Sentence**：**Raynaud phenomenon** in obliterative **arterial disease** on the upper extremity.	**SemRep**： Eicosapentaenoic Acid | ASSOCIATED_WITH | Disorder of artery **PMID**：6320945 **Sentence**：It is concluded that rheological changes that result from a diet rich in **eicosapentaenoic acid** may contribute to the suggested protective effects of such a diet against arterial disease and that such changes are of potential therapeutic importance in established **arterial disease**.

3.2.2 偏头痛与镁

Swanson 在 1988 年发现了偏头痛（Migraine）和镁（Magnesium）可能存在关联，为了能够重现他的实验，本文选取 1987 年以前的语料进行实验。本实验将初始词设为

偏头痛，进行开放式发现。再经过语义类型过滤，过滤掉了一些不相关语义类型的概念；通过 MeSH 宽泛含义概念过滤剔除了一些比较宽泛的概念，诸如：“Anions”、“Elements”等。通过目标词排序算法，最终得到的镁的排名如表 3.3，从表中可以看到镁的排名为 10，而在 Element,Ion,or Isotope（元素，离子或同位素）这一种的语义类型下排名为 1，而在 Srinivasan 实验中排名为 5。

表 3.5　偏头痛与镁的病理分析

Tab. 3.5　The pathological analysis of Migraine and Magnesium

连接词（B）	C→B 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子	B→A 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子
gamma-aminobutyric acid（γ-氨基丁酸）	**SemRep**： Migraine Disorders | ASSOCIATED_WITH | gamma-Aminobutyric Acid **PMID**：1164613 **Sentence**：The results suggest disordered **GABA** metabolism in **migraine**.	**SemRep**： Magnesium | STIMULATES | gamma-Aminobutyric Acid **PMID**：1143360 **Sentence**：At 25 degrees C the uptake of 14-C-5-HT and **14-C-GABA** was enhanced by ATP and **magnesium**.
Histamine（组胺）	**SemRep**： Histamine | ASSOCIATED_WITH | Migraine Disorders **PMID**：239368 **Sentence**：This fact may explain why, despite incriminating evidence for a pathophysiologic role of **histamine**, the usual antihistaminic agents are rather ineffective in **migrainous headaches**.	**SemRep**： Magnesium Ions | INTERACTS_WITH | Histamine **PMID**：6308958 **Sentence**：Significance of **magnesium ions** for the effects of **histamine** on contractions induced by endogenous noradrenaline.

利用 SemRep 语料对偏头痛与镁的病理分析，得到了一系列的关系组合，如表 3.5。从表中可以看到在 PMID 为 1164613 的文献中提到偏头痛患者体内的γ-氨基丁酸失常，而 PMID 为 1143360 的文献提到镁可以提高γ-氨基丁酸的含量；在 PMID 为 239368 的文献中提到偏头痛与组胺有关联，而镁离子对组胺的含量有影响。

3.2.3 老年痴呆症与消炎痛

在实验过程中，我们发现在 1993 年之后有多篇文章提及了老年痴呆症(Alzheimer's Disease)与消炎痛(Indomethacin)，为了重现二者的关系，本实验选择 1993 年之前的语料进行实验。初始词设为老年痴呆症，经过开放式发现以及过滤处理后，消炎痛的排名如表 3.3 所示。

表 3.6　老年痴呆症与消炎痛的病理分析

Tab. 3.6　The pathological analysis of Alzheimer's Disease and Indomethacin

连接词（B）	C→B 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子	B→A 在 SemRep 语料抽取的关系以及关系所在 MEDLINE 语料中的句子
Hypercapnia（血碳酸过多症）	**SemRep**： Hypercapnia | ASSOCIATED_WITH | Alzheimer's Disease **PMID**：2187699 **Sentence**：On the contrary, rest flow velocities and vasomotor responses to **hypercapnia** induced by both apnea and rebreathing tests proved to be lower in MID patients than in **SDAT** and healthy groups.	**SemRep**： Indomethacin | DISRUPTS | Hypercapnia **PMID**：6376989 **Sentence**：Both **hypercapnia** and hypoxia significantly increased CBF H/A and both increments were abolished by **indomethacin**.
Lipid Metabolism（类脂代谢作用）	**SemRep**： Alzheimer's Disease | ASSOCIATED_WITH | Lipid Metabolism **PMID**：1946550 **Sentence**：Since we have shown elevated tin levels in patients with **Alzheimer's disease**, and since organic tin compounds given to animals produce a syndrome with similarities to Alzheimer's disease, there is a need for investigation of the role of tin in **lipid metabolism** in dementia.	**SemRep**： Indomethacin | AFFECTS | Lipid Metabolism **PMID**：6809337 **Sentence**：A possible explanation is that **indomethacin** may alter chondrocyte lipid metabolism in the presence of substrate molecules by rechanneling lipid synthesis away from the prostaglandin pathway to other **lipid synthetic** pathways.

利用 SemRep 语料对老年痴呆症与消炎痛的病理分析，得到了一系列的关系组合，经过对排名靠前的目标词进行筛选，本文选取了几个有代表意义的句子和关系，如表 3.6 所示。我们可以看到在 PMID 为 2187699 的文献中提到老年痴呆症与血酸过多症（Hypercapnia）有关联，而在 PMID 为 6276989 的文献中提到消炎痛能改变血酸过多症的影响，所以我们可以得到消炎痛能够通过血酸过多症来影响老年痴呆症；在 PMID 为 1946550 的文献中提到老年痴呆症与类脂代谢（Lipid Metabolism）作用有关联，而在 PMID 为 6809337 的文献中提到消炎痛对类脂代谢作用有影响，所以我们可以得到消炎痛也可以通过改变类脂代谢作用进而影响老年痴呆症。

3.3 小结

本文提出了一种基于语义关系的非相关文献知识发现方法，旨在帮助生物医学人员发现潜在的知识。相较于以前简单 MeSH 共现的方法，通过使用 SemRep 语料，根据实

体之间的关系能够生成可解释的假设，并大幅度减少了无关的目标词数量；同时本文综合了前人在连接词和目标词过滤方面的工作，取得了不错的过滤效果；最后根据经过开放式发现后形成的概念之间的网络特点，提出了一种目标词排序算法，得到的目标词排名都相对靠前，验证了方法的有效性。

同时 SemRep 语料中抽取的实体之间的关系准确率还有进一步提升的空间（准确率为 0.73，召回率为 0.55，综合分类率 F 值为 0.63）[35]。我们下一步的计划是将命名实体识别和事件抽取方面的知识融合到实验中，以提高实体间关系的准确率、召回率，进而改善知识发现的性能。

4 基于半监督学习的生物医学文献知识发现研究

D.Hristovski 等人[17]利用 BioMedLEE 和 SemRep 两个自然语言处理系统对生物文献进行处理，对每个句子抽取语义关系，并定义了关联规则，得到了可解释的发现。但是 Semrep 的低召回率（召回率为 0.55）使得其在对句子进行处理的过程中，会使得大量的句子关系丢失,进而影响知识发现的效果。本文第三章在使用 SemRep 进行病例验证的过程中，发现许多 Swanson 使用的连接词并没有在 SemRep 出现。基于非相关文献的知识发现研究主要是为了发现潜在的知识，如果实验数据本身就不太完整，会使得发现结果大打折扣，有可能遗漏许多有用的发现。本文参考生物医学中蛋白质之间交互关系抽取的方法[36]，提出了一种半监督学习的方法，对句子中的关系进行抽取，将问题转化为句子是否存在关系的二分类问题，解决了多分类问题下召回率低的问题。

本文将假设发现经典的 ABC 模型分为 AB 和 BC 两部分，AB 模型主要是用来判断初始词和连接词是否有关系，而 BC 模型主要是用来判断连接词和目标词是否有关系。在实验过程中，通过语义类型对初始词、连接词、目标词进行限制，由于三者的语义类型都不同，进而得到两个不同的模型。由于实验的语料大部分都是非标注语料，标注语料比较少，引入半监督学习方法 Co-training 进行模型训练。Co-training 训练过程使用两个不同的分类器，要求分类器能够从不同的视图对一个问题进行描述。本文从基于词特征的核和图核两个不同的方面对句子进行表示，然后使用 Co-training 进行模型训练。实验结果表明随着训练集的不断扩充，模型的分类效果不断提高。对 Swanson 发现的经典病例进行验证，取得了较好的效果。

4.1 实验所使用工具

4.1.1 斯坦福句法分析器

上世纪九十年代，美国斯坦福大学的自然语言处理小组开发了斯坦福句法分析器（Stanford Parser）[37]，成为以后自然语言处理领域进行后续研究的基石。斯坦福句法分析器的功能主要有：一、分析句子中的句法结构，能够将句子中主谓宾结构进行提取，对于每个单词的词性进行标注。二、能够分析短语结构，诸如动宾短语、介宾短语等，不但能够直接以广义表的形式输出这些结构，还能根据这些短语结构生成句子的树形结构。可以通过句子的树形结构，对句子进行层次关系的分析，也能够利用其进行句子中各短语进行依存分析。

本文中所使用的图核特征就是在斯坦福句法分析器基础上进行抽取的，图 4.1 是一个斯坦福句法分析器关于句子“Beneficial effect of fish oil on blood viscosity in peripheral vascular disease.”的一个实例。

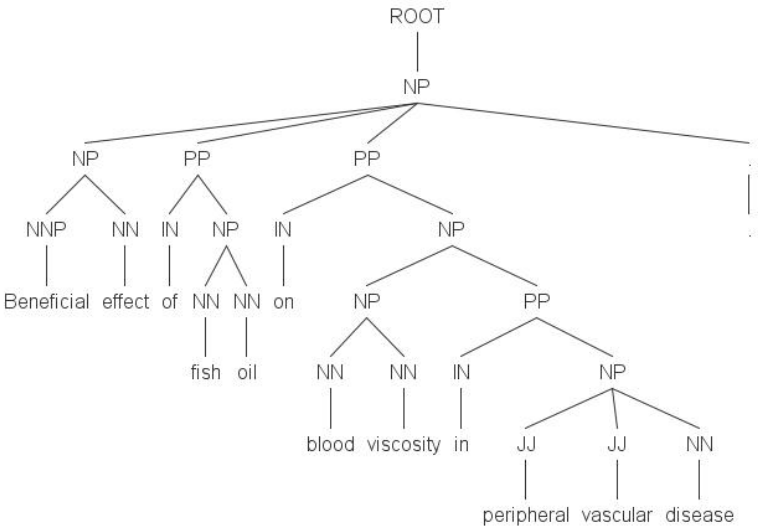

图 4.1　斯坦福句法分析器示例

Fig. 4.1　Example of Stanford Parser

4.1.2 SVM 分类器

SVM 分类器（Support Vector Machine）[38-39]是机器学习中的一种分类算法，Vapnik 等人[40]在线性分类器方面的另一重大理论创新。它的主要思想是认为可以找到待分类数据的超平面，通过这个超平面使得数据线性可分。针对线性不可分的情况，通过非线性映射的方法，对特征空间进行升维操作，将数据从低维空间映射到高维空间，进而使得数据线性可分。在数据进行升维的过程中，有可能产生维度灾难问题，Vapnik 等人引入核函数加以解决。同时由于 SVM 是基于结构风险最小化的理论，能够使得分类器的效果达到全局最优化。对于二分类问题，SVM 有很好的效果。由于本文的关系抽取实验需要进行二分类，同时通过使用有关句子关系抽取的核函数并希望结果达到全局最优，因此本文实验中所使用的分类器为 SVM 分类器。

4.2 实验方法

4.2.1 基于词特征的核

由于一个句子可以由其所包含的单词来表示，具有相似单词的句子更有可能具有相同的类别。因此，本文选择了词袋特征、概念距离特征、关系词特征、否定词特征来对句子进行描述。

本文主要使用的词特征共有如下四种：

（1）词袋特征：在一个句子中，两个概念周围的单词对两者之间是否存在关系有很大影响，本文将概念周围的单词作为词袋特征加以考虑。该特征选择句子中第一个概念左边三个单词；第二个概念右边三个单词；两个概念之间的单词。对于不同位置的单词，对它们加入不同的前缀加以区分，例如，单词“word”，在以上三种情况分别表示为l_word、r_word、m_word。在使用词袋特征时，首先对所有语料建立词典，然后对每一维特征用布尔变量表示，1 表示在句子存在此特征，0 表示在句子中不存在此特征。

（2）概念距离特征：在一个句子中，表示两个概念之间相隔多少个单词。通过分析语料特点，句子中的两个概念距离越近，两者存在关系的可能性越大。

（3）关系词特征：通过对句子文本进行分析，本文作如下假定：如果两个概念存在某种关系，那么这两个概念周围会有较大的概率出现一些动词或者它们的变体，如“activate”、“activation”、“induce”、“modulate”。通过建立关系词列表，来判断句子中是否存在关系词。关系词使用布尔变量表示，1 表示在句子存在此特征，否则为 0。

（4）否定词特征：在有些句子中，会有诸如“not”，“neither”，“no”等否定词出现，作者用来表达两个概念不存在某种关系。如果一个句子中同时出现否定词和关系词，单纯依靠关系词来判断两个概念是否有关系，会产生比较高的错误率。因此，引入布尔特征，如果有否定词出现，则置为 1；否则为 0。

4.2.2 图核

一个句子还可以通过其句法结构、主谓宾结构以及各个单词的词性进行表示，具有相同的句法结构的句子更有可能是相同的类别。我们通过将一个句子表示成图的形式，对不同的句子的图结构进行相似度计算，进而根据训练集，得到待分类数据的类别[41]。

图核的原理是通过斯坦福句法分析器对句子进行分析，得到句子的依存关系和词性信息。根据这些信息，将句子用带权有向图来表示，包含两个不相连接的子图：句子的依存结构子图（PSS）；句子中词之间的线性序列子图（LOS）。Airola等人提出的全路径图核[42]主要使用了以上两个子图。

图核计算句子间相似度的方法是通过比较不通句子之间的顶点标签关系来实现的。在图中，每个词和依存关系都会产生各自的顶点，这些顶点都用句子中的词和词性以及标签来表示。词和词性组成的顶点的表示形式，如图4.2中PROT1_IP NN_IP，其中PROT1_IP和NN_IP；依存关系组成的顶点，如图中的nsubj。在最短路径子图中，所有顶点都有各自的标签IP；而在线性顺序子图中有(B)efore, (M)iddle, and (A)fter三种类型的标签。

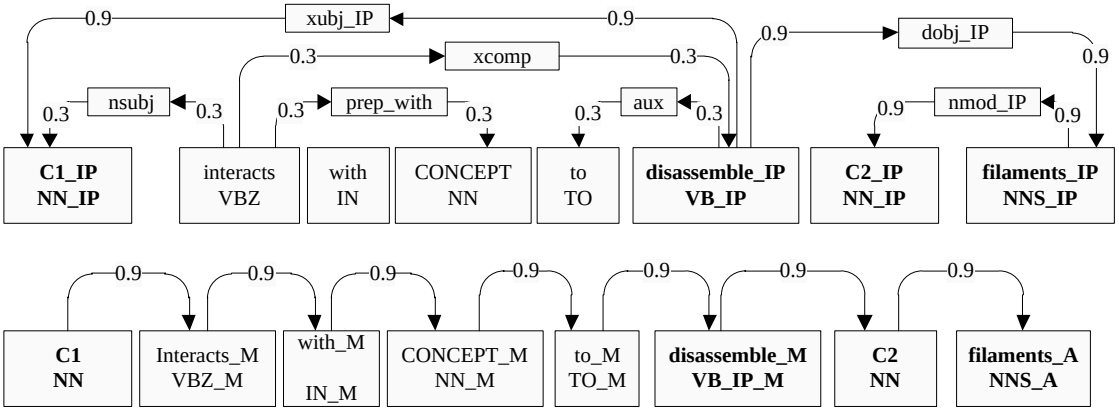

图 4. 2　图核的一个表达实例

Fig.4.2　An example of Graph Kernel

在一个句法结构中，可以通过判断两个概念之间的词或者连接它们的词来判断概念之间是否存在关系。在句子依存结构子图中，连接两个概念路径上的节点是不同的，给每条路径上的边分配不同的权值，最短路径上的边的权值为0.9，否则为0.3。在线性序列子图中，每个词根据其B（bettwen），M（middle）和A（after）标签决定其在两个概念的前后位置。通过一条边连接前面后相邻的两个词，边的权值设置为0.9。

例如，图4.2是一个图核表达示例，通过分析句子中两个概念C1、C2在句子中的结构，对不同的边进行赋值。图4.2的上半部分为句子的依存结构子图（PSS）；其下半部分为句子中词之间的线性序列子图（LOS）。

在图核中，计算句子之间相似度的方法主要根据PSS和LOS两个子图信息，通过计算不同句子之间图的相似度来得到不同句子的相似度。而子图的信息主要是顶点的标签关系，使用图矩阵G来表示一个子图，具体计算方式如公式4.1所示：

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

其中，M的行和列表示的是各个顶点，它是一个图的邻接矩阵。M_{ij}表示连接顶点V_i和V_j之间边的权重。L是标签矩阵，$L_{ij}=1$ 表示顶点j中包含第i个标签；而$L_{ij}=0$表示顶点j不包含。

由此，我们定义两个图矩阵G和G'之间的图核k(G , G')计算方法，如公式4.2所示：

$$k(G,G^{'})=\sum_{i=1}^{L}\sum_{j=1}^{L}G_{ij}G_{ij}^{'} \qquad (4.2)$$

4.2.3 Co-training

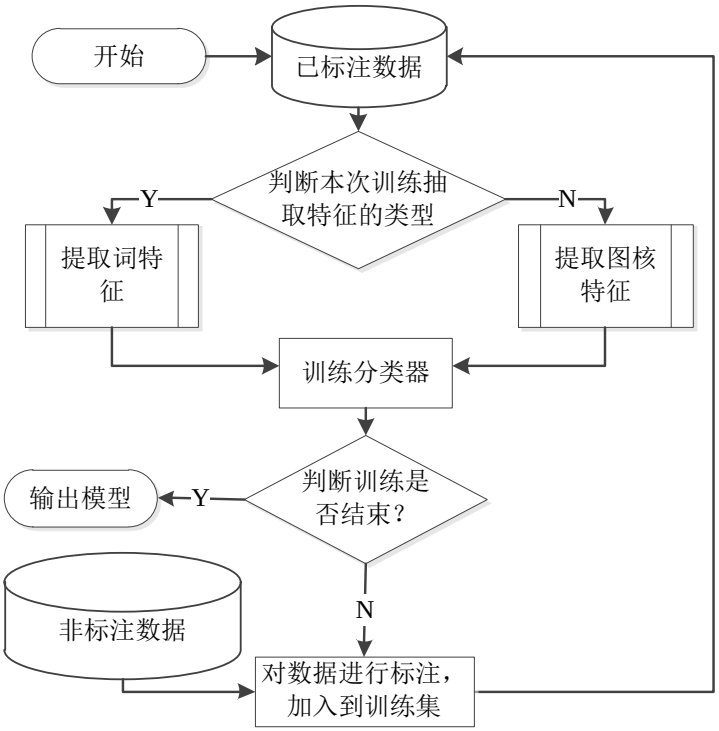

图 4.3　Co-training 训练流程图

Fig.4.3　The flow chart of Co-training

在进行文本挖掘的过程中，经常会遇到语料大部分都是未标注数据，只有少部分为标注数据的情况。而直接使用少量的未标注数据进行模型训练，往往不能得到好的实验结果。Co-training[43-46]是一种半监督的机器学习方法，是 bootstrapping 思想的一种实现。它的基本思想是：首先建立两个不同的分类器 C_1、C_2，利用 C_1 对标注的语料进行训练得到初始分类器，然后对未标注语料进行训练，从训练结果中选择可信度大的实例，并将这些实例加入到训练集中；然后使用 C_2 对更新后的训练集进行训练，得到新的分类

器，继续对未标注语料进行训练，并选择可信度大的实例加入到训练集。如此交替进行训练，直到没有未标注语料可供使用或者模型不再变化为止。

同一个数据可以从多种不同的角度进行描述，即是数据的不同视图。比如，在本实验中，一个句子可以通过句子的词特征对句子进行表示，也可以通过图核特征来对句子进行表示。通过不同视图建立的两个分类器交替训练，能够避免自训练（Self-Training）算法的缺点：随着训练集不断增大，训练模型的错误也不断累积，最终导致模型结果效果不佳。

由于 Co-training 并不能剔除训练集已有的噪音，在 Co-training 训练过程中，不断地从未标注语料挑选实例加入到训练集中，如果对未标注数据选择的方法不好，就会导致训练模型的性能非但不能提高，反而积累了噪音。所以未标注数据的选择标准是决定 Co-training 训练能否有效的关键。本文根据 SVM 分类器的特点，使用 SVM 分类器对未标注数据的打分作为评价标准。在 SVM 分类器进行二分类问题时，如果一个数据的结果大于本文设定的阈值，就判定为正例；否则，为负例。而一个数据得分与阈值差值的绝对值越大，便认为其离分类超平面的距离越大，这个数据被误分类的可能性也越低。我们分别设置每次添加未标注数据到训练集占总未标注数据的比例为 0.05，0.1，0.15，0.2 进行实验。通过实验表明，AB 模型，本文选择每次添加的比例为 0.1；BC 模型，本文选择每次添加的比例为 0.15。如果比例值选择较大，每次加入的噪音会变多，使得分类效果降低；而如果比例值选择较小，训练集的大小增长会比较缓慢，模型改善不明显。

在本实验中，具体的训练过程如图 4.3 所示，通过交替抽取不同的特征，训练不同的分类器，对未标注语料进行预测，将预测的结果加入到训练集中，达到扩充训练集的目的，当训练结束后，输出训练模型。

4.2.4 语义类型过滤

本文使用的语义类型与第三章的语义类型过滤一致，连接词、目标词的语义类型如表 3.1 所示，而初始词的语义类型，本文设置为“Disease or Syndrome”（疾病与症状）。通过对三者进行语义类型限制，AB 和 BC 两个模型有了不同的作用。AB 模型主要用来预测一种疾病是否存在或者引起一种生理现象，BC 模型主要用来预测一种物质是否能够引起一种生理现象。这样，通过结合 AB、BC 这两个不同的模型，得到完整的 ABC 模型，进而进行假设发现研究。

4.3 实验结果及分析

本文所使用的语料来源为 SemRep 数据库和 MEDLINE 数据库，主要使用 SemRep 数据库提供的带有实体之间语义关系的句子，这主要是因为 SemRep 已经对句子做过处

理，在 SemRep 数据库中的句子存在关系的可能性比较高。但是 Semrep 语料中识别出的两个概念是标准化的概念，无法将概念和原始句子进行匹配。本文使用 MetaMap 对原始句子进行命名实体识别，将概念与原始句子中的单词进行映射。得到数据后，通过对这些句子进行手工标注，形成实验所用的初始训练集和测试集。在对 Swanson 发现的经典病例进行验证的过程中，本文使用 MEDLINE 数据库进行检索，最大限度的保留连接词和目标词。

在 AB 和 BC 两个模型中，本文各自分别建立了两个不同的分类器，一个是使用基于词特征的核建立的，一个是基于图核特征建立的。基于词特征的分类器使用训练集建立模型的主要过程如下：首先，对包括训练集和测试集在内的所有句子建立词典；其次，使用词典结合规则的方法对训练集中的词特征进行抽取，将训练集中的实例用向量空间的方式表示出来；然后，对使用词特征表示的句子利用 SVM 分类器进行训练，得到分类器模型；最后，使用训练后的模型对测试集合未标注语料进行预测。基于图核特征的分类器使用训练集建立模型的主要过程如下：首先，对句子使用斯坦福句法分析器进行句法解析；其次，对解析完的句子使用图核提取特征的方法进行特征抽取；然后，使用 SVM 分类器对训练数据进行模型训练；最后，使用训练后的模型对测试集和未标注语料进行预测。

4.3.1 实验评价指标

1. 准确率、召回率、F 值

在本文中，建立了两个模型 AB、BC 模型，对两者都进行了 Co-training 训练。实验所使用的评价指标主要有准确率（P）、召回率（R）、F 值（F），三者的计算方式如公式 4.3、4.4、4.5 所示：

$$P = \frac{TP}{FP} \tag{4.3}$$

$$R = \frac{TP}{FN} \tag{4.4}$$

$$F = \frac{2 * P * R}{P + R} \tag{4.5}$$

其中，*TP* 为预测正确的数据个数，*FP* 为所有预测数据的个数，*FN* 为测试集中所有正例的个数。*P* 是用来评价模型的准确率。*R* 是用来评价模型的召回率，*F* 值是准确率和召回率的调和平均数。本实验中，主要使用这三个指标对模型结果进行评价。

2. 有效连接词

本文定义能够将初始词和目标词连接起来的连接词为有效连接词，有效连接词占总连接词的比例称为有效连接词的比例，具体的计算方法如公式 4.6 所示：

$$S = \frac{n}{N} \tag{4.6}$$

其中，n 为有效连接词的数量，N 为与初始词有关联概念的数量。

4.3.2 AB 模型

本模型主要考察初始词和连接词是否存在关系，也即是判断一种疾病是否会存在或者导致一种生理现象的产生。

本文从医学主题词 MeSH 中选择了 200 个语义类型为“Disease or Syndrome（疾病与症状）”的概念，使用这 200 个概念对 Semantic MEDLINE Database[32]进行检索，得到实验所使用的句子。经过 MetaMap 进行命名实体识别处理后，经过语义类型过滤，限制初始词和连接词的语义类型，一共得到了 20395 个句子。通过对句子进行手工标注，形成实验的初始训练集和测试集，它们的大小分别为 600 个和 500 个。在进行人工标注时，本实验所使用的标注标准如下：如果一个句子中的两个概念含有关系词列表中的关系，便认定两者存在正关联，标为正例，而诸如“B in A.”、“A can change B.”等句子本实验也标为正例，因为“B in A”，说明 B 这种生理现象在 A 这种疾病中；如果 A 和 B 两个概念在句子中没有明显的关系仅仅只是共现则标为负例，而诸如“A is a B.”、“A and B”等句子本实验也标为负例，因为“A is a B”是一个上下级“ISA”的关系，不是本实验所需要的关系，“A and B”仅仅只是 A 和 B 共现，并不能表示 A 能够引起 B 的变化。

本实验首先使用图核特征对初始训练集 T_1 进行模型训练，得到模型 M_1，使用 M_1 对测试集进行测试，通过对实验阈值进行调整，我们选择最佳结果的阈值 t=0.0425，得到的结果为：P=72.88%，R=83.33%，F=77.76%。使用模型 M_1 对 2000 个未标注句子进行预测，分别选择得分最高和最低的 10%的句子，即正负实例各 200 个，加入到训练集中，得到新的训练集 T_2。然后使用基于词特征的核对训练集 T_2 进行训练，得到模型 M_2，使用 M_2 对测试集进行测试，通过对实验阈值进行调整，选择能够得到最佳结果的阈值 t =-0.1307，得到的结果为：P=74.35%，R= 76.33%, F=75.33%。通过 Co-training 对模型进行交替训练，不断扩充训练集。

随着训练集的扩大，每次处理的未标注语料的也不断扩大。实验训练集一共进行了 5 次扩充，每次处理的未标注句子的规模依次为 2000，3000，4000，5000，6000，训练集的规模也从最开始的 500，依次增加为 900，1500，2300，3300，4500。在基于特征

的核和基于图核的模型上，我们分别对这 5 次训练集的变化在测试集上进行测试，得到它们各自的准确率（P），召回率（R），F 值（F），具体的结果如图 4.4、图 4.5 所示。

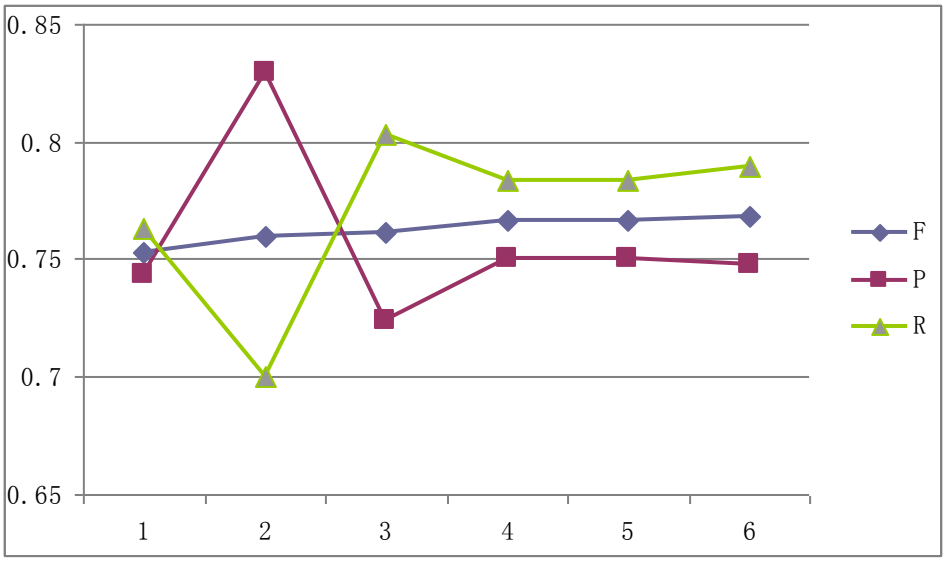

图 4.4　AB 模型在特征核的结果

Fig.4.4　The result of AB model based on Bag of Words

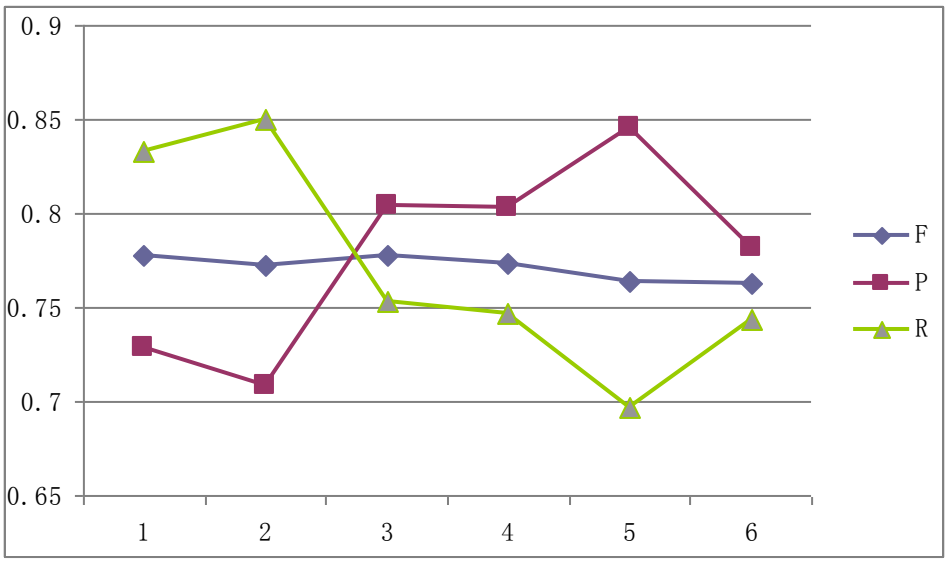

图 4.5　AB 模型在图核的结果

Fig.4.5　The result of AB model based on kernel of graph

从图 4.4 中可以看到，在经过了五次训练集扩充后，虽然基于特征核的分类器的准确率、召回率不断变化，但 F 值在不断提高，最终的准确率为 74.76%，召回率为 79%，F 值为 76.82%。

从图 4.5 中可以看到，基于图核的分类器效果并没有得到太多提高，其最高值出现在第二次训练集扩充后，准确率为 80.43%，召回率为 75.33%，F 值为 77.8%。

两个模型的最好结果都比使用初始训练集训练的模型效果好，所以在句子中概念之间的关系抽取中，使用半监督学习方法 Co-training 改善模型的效果是可行的。同时，通过对比两个使用不同核的模型，图核模型的结果比特征核模型的结果效果好，所以本文在使用本方法进行假设发现有效性验证的时候，选择图核模型进行实验。

4.3.3 BC 模型

本模型主要考察连接词和目标词之间的关系，也即是判断一种物质能否改变人的生理现象或者生理过程。

老年痴呆症（Alzheimer's Disease）和镁（Magnesium）作为 Swanson 发现的经典病例之一，我们使用老年痴呆症检索 SemRep 数据库得到的连接词，共计 346 个（已使用语义类型对其进行过滤），利用开放式发现算法，再次对 SemRep 数据库进行检索，得到有关连接词和目标词的句子。通过使用 MetaMap 进行命名实体识别和语义类型过滤，得到实验数据，一共得到了 20490 个句子。类似于 AB 模型，我们也对这些句子进行手工标注，形成实验所使用的初始训练集和测试集，它们的大小分别为 800 个和 500 个。但在进行人工标注时，所使用的标注标准与 AB 模型有不同，具体如下：如果一个句子中的两个概念含有关系词列表中的关系，我们便认定两者存在正关联，标为正例，而类似于“B in C.”这种类型的句子标为负例，因为在 BC 模型中，我们需要的是 C 能够改变 B 的关系；对于其他负例的标注，类似于 AB 模型的标注。

BC 模型的训练过程与 AB 模型类似，我们首先使用图核特征对初始训练集 T_1 进行模型训练，得到模型 M_1，使用 M_1 对测试集进行测试，通过对实验阈值进行调整，我们选择最佳结果的阈值 t=-0.3365，得到的结果为：P=78.76%，R=95.08%，F=86.15%。在使用模型 M_1 对 2000 个未标注句子进行预测时，我们分别选择得分最高和最低的 15%的句子，即正负实例各 300 个，加入到训练集中，得到新的训练集 T_2。然后再使用基于特征的核进行模型训练，进行 Co-training 训练过程。

BC 模型的训练集一共也进行了 5 次扩充，每次处理的未标注句子的规模依次为 2000，3000，4000，5000，4827，训练集的规模也从最开始的 500，依次增加为 1100，2000，3200，4100，6017。类似于 AB 模型，我们也分别在两个特征核针对 5 次训练集

变化，在测试集上进行测试，得到各自的准确率（P），召回率（R），F 值（F）的变化趋势，如图 4.6、图 4.7 所示。

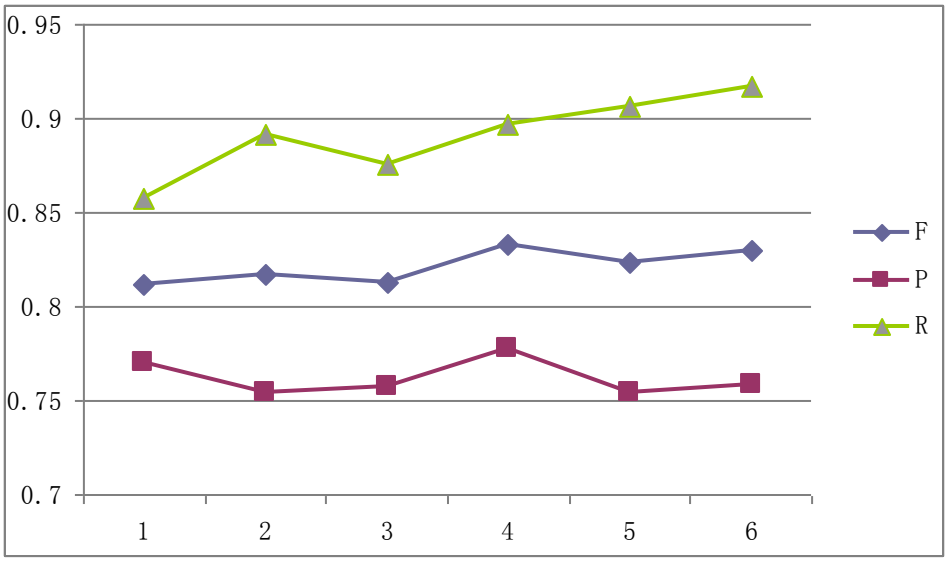

图 4.6　BC 模型在特征核的结果

Fig.4.6　The result of BC model based on Bag of Words

在图 4.6 中，　BC 模型在 5 次训练集扩充后，基于词特征核分类器的 F 值达到了最高值，准确率为 75.8%，召回率为 91.71%，F 值为 83%。

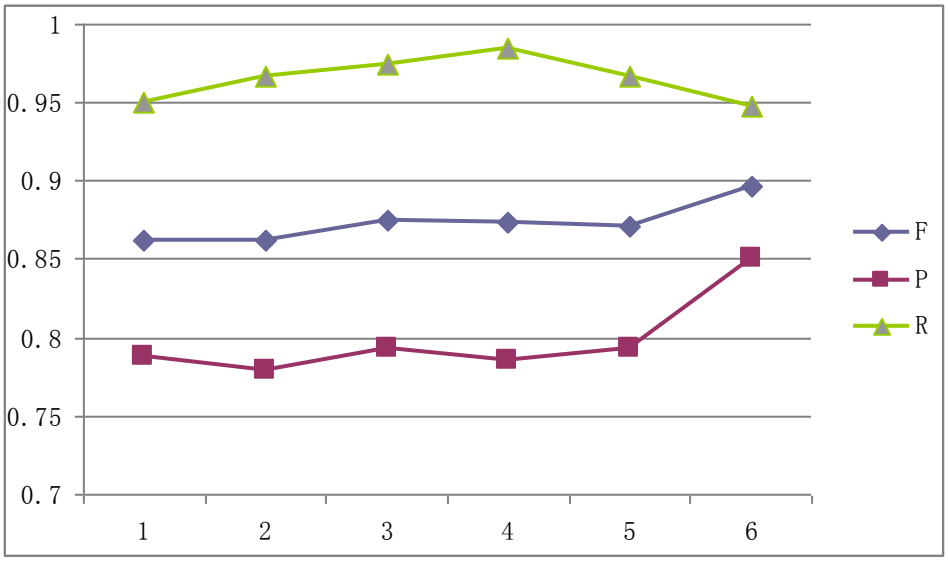

图 4.7　BC 模型在图核的结果

Fig.4.7　The result of BC model based on kernel of graph

从图 4.7 中我们看到，基于图核模型的 F 值在最后一次训练集扩充后达到了最高值，准确率为 85.13%，召回率为 94.82%，F 值为 89.71%。

在 BC 模型中，两个模型的最好结果也都比使用初始训练集训练的模型效果好，进一步证明了半监督学习方法 Co-training 在句子中概念之间的关系抽取中的有效性。 同时，在 BC 模型中对比两个使用不同核的模型，图核模型的结果比特征核模型的结果效果好，所以本文在使用本方法进行假设发现有效性验证的时候，选择图核模型进行实验。

4.3.4 实验方法在假设发现的应用

表 4.1　有效连接词数目以及比例

Tab.4.1　The number and scale of the useful linking concept

病例	实验方法	连接词数目	有效连接词数目	有效连接词比例
雷诺氏病与鱼油	本文所提方法	255	21	8.2%
	SemRep	132	2	1.5%
偏头痛与镁	本文所提方法	620	294	47.41%
	SemRep	491	78	15.88%

本文使用得到的 AB 和 BC 模型对 Swanson 发现的雷诺氏病和偏头痛进行验证，并与使用 SemRep 工具进行知识发现研究产生的结果进行对比。

首先，对雷诺氏病使用闭合式发现算法，将检索词分别设置为 Raynaud Disease（雷诺氏病）和 Fish Oils（鱼油），检索 1986 年以前的 MEDLINE 语料。其次，通过 MetaMap 进行命名实体识别，得到包含雷诺氏病或者鱼油的句子。然后，利用语义类型过滤对句子中不符合的句子进行过滤。最后使用 AB 模型对包含雷诺氏病的句子进行分类，使用 BC 模型对包含鱼油的句子进行分类。对分类结果为正例的句子中的连接词进行提取，找到有效连接词。本文使用 SemRep 数据库进行知识发现研究做为对比实验。具体的实验过程为使用 Raynaud Disease（雷诺氏病）和 Fish Oils（鱼油）做为检索词，对 1986 年之前的 SemRep 数据库进行检索，对概念进行语义类型和语义关系过滤后，采用闭合式算法将能够连接雷诺氏病和鱼油的有效连接词统计出来。同时，本文也对偏头痛和镁的案例进行了验证，具体的实验结果如表 4.1 所示。

从表 4.1 中可以看到，本文提出的方法相较于 SemRep 在有效连接词的数量和比例方面都有很大的提高，能够发现更多的连接疾病和治疗方法的通路，也即能够发现更多的潜在信息。由于知识发现的目的就是发现已有文献的潜在信息，所以能够找到更多的有效连接词是改善知识发现方法的有效途径。

4.4 小结

本文使用基于语义关系的方法进行知识发现研究，着重改进 SemRep 工具的召回率不高问题，通过使用半监督学习方法结合词特征核和图核进行句子中概念之间关系的抽取，在句子关系抽取方面相较于 SemRep 有提高。本文对知识发现的开放式算法进行分析，将 ABC 模型分为两个部分，并使用 SVM 分类器分别建立了 AB 和 BC 两个模型，综合两个模型进行实验得到可解释的发现。最后与 SemRep 工具在经典病例上进行对比，实验结果表明该方法能够发现更多的有效连接词，能够发现更多的潜在信息。

结　论

本文主要对基于生物医学文献的知识发现进行了研究，综合使用了生物医学资源和机器学习的知识，提出了基于语义和基于非监督学习的两种方法，并对这两种方法进行了论证和分析。

本文提出了一种基于语义资源的方法，旨在综合现有的假设发现工具，结合自己提出的排序算法进行开放式假设发现研究。首先利用 SemRep 工具抽取句子内实体之间的关系；其次结合语义类型、概念的信息量以及关联规则对连接词、目标词进行过滤；最后根据统计量信息对目标词进行排序。本文的方法在 Swanson 发现的经典病例进行了验证。实验结果表明，该方法不但能够使得有效的目标词排名靠前，而且还能够对发现的知识进行解释。

在基于非监督学习的方法中，本文着重改进 SemRep 工具的召回率不高问题。本文参考从句子中抽取蛋白质交互关系的方法，利用半监督学习方法抽取句子中概念之间的交互关系。本文使用特征核和图核抽取句子之间的关系，并使用 Co-training 进行训练集的扩充，在句子关系抽取方面相较于 SemRep 有提高。本文使用以上的关系抽取方法，利用 SVM 分类器，分别建立了 AB、BC 两个模型，对于不同语义类型的关系分别进行抽取，并与 SemRep 工具在经典病例上进行对比。实验结果表明该方法取得较好的效果。

在今后的研究中，使用机器学习的方法从语义角度对连接词、目标词进行过滤仍然会是生物医学文献知识发现的主流。以后的工作可以从以下两个方面展开，一方面，扩大知识发现的语料集合，引用 DNA 微阵列和高通量信息等数据，通过扩大学科之间的交叉，发现有价值的假设；另一方面，将知识发现的理论应用到药物副作用发现，药物重定位等领域，发挥知识发现理论的实用价值。

参 考 文 献

[1] National Library of Medicine. MEDLINE Fact Sheet[OL].[2010-04-25]. http://www.nlm.nih.gov/pubs/factshee ts/MEDLINE.html.

[2] Swanson D R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge[J]. 1986.

[3] Swanson D R. Undiscovered public knowledge[J]. The Library Quarterly, 1986: 103-118.

[4] Swanson D R. Migraine and magnesium: eleven neglected connections[J]. Perspectives in biology and medicine, 1987, 31(4): 526-557.

[5] Swanson D R. Two medical literatures that are logically but not bibliographically connected[J]. Journal of the American Society for Information Science, 1987, 38(4): 228-233.

[6] Swanson D R. Medical literature as a potential source of new knowledge[J]. Bulletin of the Medical Library Association, 1990, 78(1): 29.

[7] Swanson D R. Somatomedin C and arginine: implicit connections between mutually isolated literatures[J]. Perspectives in biology and medicine, 1990, 33(2): 157.

[8] Smalheiser N R, Swanson D R. Using Arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses[J]. Computer methods and programs in biomedicine, 1998, 57(3): 149-153.

[9] Gordon M D, Lindsay R K. Toward discovery support systems: A replication, re‐examination, and extension of Swanson's work on literature‐based discovery of a connection between Raynaud's and fish oil[J]. Journal of the American Society for Information Science, 1996, 47(2): 116-128.

[10] Weeber M, Klein H, de Jong‐van den Berg L, et al. Using concepts in literature‐based discovery: Simulating Swanson's Raynaud‐fish oil and migraine‐magnesium discoveries[J]. Journal of the American Society for Information Science and Technology, 2001, 52(7): 548-557.

[11] Weeber M, Vos R, Klein H, et al. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide[J]. Journal of the American Medical Informatics Association, 2003, 10(3): 252-259.

[12] Srinivasan P. Text mining: generating hypotheses from MEDLINE[J]. Journal of the American Society for Information Science and Technology, 2004, 55(5): 396-413.

[13] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery[J]. Journal of biomedical informatics, 2006, 39(6): 600-611.

[14]Hu X. Mining Novel Connections from Large Online Digital Library Using Biomedical Ontologies [J]. Library Management Journal,2005,26(4):261-270.

[15]Hu X, Xiaodan Zhang, lllhoi Yoo, et al. Mining Hidden Connections Among Biomedical Concepts from Disjoint Biomedical Literature Sets Through Semantic-Based Association Rule [J]. International Journal of Intelligent Systems,2010,25:207-223.

[16]Miyanishi T, Seki K, Uehara K. Hypothesis Generation and Ranking Based on Event Similarities [C]. Proceedings of the 2010 ACM Symposium on Applied Computing, Sierre, Switzerland,2010:22-26.

[17] Hristovski D, Friedman C, Rindflesch T C, et al. Exploiting semantic relations for literature-based discovery[C]//AMIA annual symposium proceedings. American Medical Informatics Association, 2006, 2006: 349.

[18] Cohen T, Widdows D, Schvaneveldt R W, et al. Discovery at a distance: Farther journeys in predication space[C]//Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on. IEEE, 2012: 218-225.

[19] Cameron D, Bodenreider O, Yalamanchili H, et al. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications[J]. Journal of biomedical informatics, 2013, 46(2): 238-251.

[20]Hristovski D, Kastrin A, Peterlin B, et al. Combining semantic relations and DNA microarray data for novel hypotheses generation[M]//Linking literature, information, and knowledge for biology. Springer Berlin Heidelberg, 2010: 53-61.

[21]刘耀, 段慧明, 穗志方. 非相关文献知识发现的数据基础研究——以中医药古文献语言知识库的构建为例[J]. 情报杂志, 2006, 25(9): 104-107.

[22]许建阳, 马明, 王梅康, 等. Swanson 的非相关文献知识发现法对中医学发展的启示[J]. 世界科学技术: 中医药现代化, 2005, 7(3): 48-52.

[23]李文林, 葛月兰, 宿树兰, 等. 利用知识发现工具 Arrowsmith 探讨当归与痛经的相关性[J]. 中华医学图书情报杂志, 2008, 17(4): 7-11.

[24]曹志杰, 冷伏海. 非相关文献知识发现方法在航天科技情报研究中的应用分析[J]. 情报理论与实践, 2008, 31(4): 569-572.

[25] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(suppl 1): D267-D270.

[26] Lipscomb C E. Medical subject headings (MeSH)[J]. Bulletin of the Medical Library Association, 2000, 88(3): 265.

[27] Aronson A R. The MetaMap mapping algorithm[R]. Internal report, 2000.

[28] Aronson A R. MetaMap Variant Generation[J]. 2007-08-01]. http://skr, nlm. nih. gov/papers/references/mm, variants. pdf, 2001.

[29] Aronson A R. MetaMap Candidate Retrieval[J]. Semantic Knowledge Representation Group, 2001.

[30] Aronson A R. MetaMap Evaluation[J]. Unpublished manuscript, 2001.

[31] Cutting D, Kupiec J, Pedersen J, et al. A practical part-of-speech tagger[C]//Proceedings of the third conference on Applied natural language processing. Association for Computational Linguistics, 1992: 133-140.

[32] National Library of Medicine, Semantic MEDLINE Database,http://skr3.nlm.nih.gov/SemMedDB/

[33] Stegmann J, Grohmann G. Transitive text mining for information extraction and hypothesis generation[J]. arXiv preprint cs/0509020, 2005.

[34] Seco N, Veale T, Hayes T. An intrinsic information content metric for semantic similarity in wordnet [C]. Proceedings of European Conference on Artificial Intelligence, Valencia,Spain,2004:1089-1090.

[35] Ahlers CB, Fiszman M, Demner-Fushman D, Lang F, Rindflesch TC. Extracting semantic predications from MEDLINE citations for pharmacogenomics. Pac Symp Biocomput 2007;2006:209 - 20.

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

[37] Klein D, Manning C D. Accurate unlexicalized parsing[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 423-430.

[38] Cortes C, Vapnik V. Support vector machine[J]. Machine learning, 1995, 20(3): 273-297.

[39] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3): 293-300.

[40] Vapnik V. The Nature of 6tatistical Learning Theory[J]. Data Mining and Knowledge Discovery, 6: 1-47.

[41] Borgwardt K M, Kriegel H P. Shortest-path kernels on graphs[C]//Data Mining, Fifth IEEE International Conference on. IEEE, 2005: 8 pp.

[42] 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.

[43] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.

[44] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training[C]//Proceedings of the ninth international conference on Information and knowledge management. ACM, 2000: 86-93.

[45] Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets[C]//Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. 2001: 1-9.

[46] Mihalcea R. Co-training and self-training for word sense disambiguation[C]//Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004). 2004.

攻读硕士学位期间发表学术论文情况

1　基于语义资源的生物医学文献知识发现研究. **李宗耀**，杨志豪. 中文信息学报（录用）. 主办单位：中国中文信息学会，中国科学院软件研究所。中文核心期刊，国内刊号 CN11-2325/N。（本硕士学位论文第三章）

致　　谢

三年的硕士研究生生活就要落下帷幕，回首这段日子，有太多的回忆与不舍，有欢乐，有忧愁，在此要感谢那些给与我帮助的老师与同学，是你们让我的这段生活多姿多彩，充满了乐趣。

首先，我要由衷的感谢我的导师杨志豪老师。在学术研究上，杨老师以严谨负责的态度，深深地影响了我。在自己研究遇到困难的时候总是能帮我找出问题的关键所在，给我提出指导性意见。在生活中，杨老师为人正直、谦虚诚恳，教会了我很多做人做事的道理，这些都成了我以后为人处世的行为准则。

其次，要感谢林鸿飞老师，作为教研室的带头人，总是尽自己最大的努力让我们的科研环境变得不枯燥。每年组织我们进行各种各样的活动，滨海路徒步走、新年元旦晚会、羽毛球比赛、文学讲座等等。通过这些活动，使得我们切身的感受到自己是在信息检索实验室这个大家庭中，这里是我们第二个家。

然后，还要感谢王健老师，王老师在面对新的研究课题时努力认真的态度令我印象深刻，王老师待人和蔼可亲，不管在科研上还是生活上，都积极的给我们提供帮助。

最后，我要感谢我的同学和师兄师姐师弟师妹，是因为你们，我的这三年才充满欢乐。在科研上，我们相互探讨、论证问题的正确性；在生活中，我们相互理解、相互帮助，倾诉各自的迷茫与困惑，勇敢地大步向前。

向所有关心我帮助过我的人表达我深深的谢意！

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作者签名：_____ 日期：_____年____月____日

导师签名：_____ 日期：_____年____月____日