

**学术学位硕士研究生学位论文中期报告**

论文题目：基于深度学习与语义分析的药物不良反应发现

|  |  |
| --- | --- |
| 姓 名： | 郭凯 |
| 学 号： | 21409162 |
| 学科专业： | 计算机技术 |
| 指导教师： | 林鸿飞 教授 |
| 入学日期： | 2014年9月1日 |
| 报告日期： | 2016年10月12日 |
| 报告地点： | 大连理工大学创新园大厦 |

研究生院制表

说 明

硕士学位论文中期检查是保证学位论文质量、工作进度和研究生培养质量的重要措施。原则上，要求硕士生应在第4学期末进行中期检查，其中，2年制专业学位硕士生应在第3学期末进行中期检查。

一、考核内容：学位论文内容完成情况、阶段性成果是否正确，开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等。

二、考核时间：原则上，硕士生的中期检查应在第4学期末进行，其中两年制专业学位硕士生的中期检查可在第3学期末进行。

三、报告撰写：中期报告正文字数不少于4000字，正文及参考文献等撰写要求参见《大连理工大学博士学位论文格式规范》。

四、考核办法：由学部（学院）组织，3-5名本学科领域硕导及以上专家组成评审专家组以答辩的方式进行。学生进行口头陈述时间不得少于10分钟。专家组给出考核成绩和是否通过的意见。

五、报告保存：中期报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：中期考核后，学生应及时登录研究生信息管理系统上传中期报告（PDF文档）及考核结果。

**硕士学位论文中期检查报告正文**

撰写大纲：

1. 开题时拟定的研究方案、进度计划；若开题时的研究方案已经调整，应说明调整的原因、调整后该领域的国内外研究状况分析、研究内容、研究方法、进度计划等；
2. 学位论文的研究进展完成情况、阶段性成果和创新点论述；
3. 后续工作的设想、可能遇到的困难和问题及条件保障措施；
4. 已发表、录用的论文和已投稿的论文情况。
5. 参考文献（不占字数）。

目 录

[1 绪论 1](#_Toc439664717)

[1.1 研究方案 1](#_Toc439664718)

[1.2 进度计划 1](#_Toc439664719)

[2 研究内容概述 6](#_Toc439664721)

[2.1 研究进展与阶段性成果 6](#_Toc439664722)

[2.2 创新点论述 8](#_Toc439664724)

[3 后续工作展望 6](#_Toc439664721)

[3.1 后续工作的设想 6](#_Toc439664722)

[3.2 可能遇到的困难和问题 8](#_Toc439664723)

[3.3 条件保障措施 8](#_Toc439664724)

[4 已发表、录用的论文和已投稿的论文情况 13](#_Toc439664726)

[参考文献 17](#_Toc439664730)

# 1 绪论

* 1. 研究方案

药物不良反应是患者在使用某种药物治疗疾病的时候产生的与治疗无关的作用，而这种作用一般都对患者不利。鉴于药物不良反应对公众安全的严重威胁，各国都建立了严格的药物不良反应申报程序，为安全用药收集了重要的基本数据，相比于上市前的临床实验等传统手段，基于此类数据的不良反应挖掘手段可以打破时间成本与金钱成本的限制，大大加速了不良反应的发现进程，具有重要的意义。

本文采取的研究方案是：本文所采用的数据集为美国食品药品监督管理局（FDA）提供的药物不良事件报告[1]，首先对数据进行清洗、利用SIDER数据库[2]和CHV数据库[3]提取药名与不良反应的实体；再使用关联规则过滤数据，并提取药物与不良反应数据对及其相应权重值，生成相应短句集合作为分布式向量训练的输入数据，计算药物与不良反应特征向量计算相应的关联度，并利用已有的生物医学数据库进行结果过滤，得到潜在的药物与不良反应关系。研究框架图如图1.1所示。整个系统主要包括不良反应报告获取、药物与不良反应实体识别与标准化、使用关联规则挖掘药物与不良反应关系、生成短句集合、生成分布式向量、计算关联度等模块。

1. **药名文本识别与标准化**

OpenFDA以JSON格式提供药物不良事件报告，报告中包含了用药信息以及不良反应报告的相关信息（如收集时间、来源渠道等）本文提取报告中“病人域”中的数据作为实验初始数据，该域中包含了“药名域”与“不良反应域”等信息，但是“药名域”仍然是非结构化的数据，具有大量的不规范写法等影响实验准确度的文本噪音，因此本文先利用文本处理的相关技术对药名文本进行预处理，为下一步的数据做准备，具体内容包括：清洗药物列表文本中的药品用量信息，过滤\*+&/等非英文字母符号，去除()[]//中包含的内容等。处理结果如表1所示。



图 1.1系统流程图

表 1 药名文本预处理

Tab 1 Pre-processing of drug text

|  |  |
| --- | --- |
| **药名文本** | **预处理结果** |
| LUPRON DEPOT 3.75 MG | LUPRON DEPOT |
| NOVORAPID (INSULIN ASPART) | NOVORAPID |
| INDOMETHACIN /00003801/ | INDOMETHACIN |
| HYDROCODONE/APAP | HYDROCODONE APAP |

经过药名文本清洗后，药名区域仍存在非药品实体，如“food supplements”,“vitamin C”等，这些都不是实验范围内的药物名称。因此本文需要识别出药名文本中的药物名称实体。

MetaMap[4]是由National Library of Medicine（NLM）开发的，用于将生物医学文本映射到UMLS超级叙词表的工具，同时也可以直接返回该生物医学文本对应的标准名称与语义类型，本文使用MetaMap对语料中药名区域中的所有名词进行映射，得到每个名词的标准名称与语义类型。

MetaMap共能识别出133种语义类型，但其中只有22种可以作为药物的语义类型，对MetaMap的药名识别结果进行过滤，完成药名识别与标准化。

1. **基于关联规则的数据处理**

从FDA报告官方文档中可知，不良反应报告中包含多个药物与不良反应时，药物与不良反应之间对应关系是不确定的，无法判断某一药物某一不良反应之间是否有联系，这就给接下来潜在不良反应挖掘增加了难度。如图2就展示了一幅不良反应事件报告例子，该报告中包含了3种药物与4种不良反应。

图 2不良事件报告举例

 本文将每个不良反应报告分割为小的关系对，每个关系对由一个药物A和不良反应B组成，如图2的不良反应报告为例，3个药物与4个不良反应可以生成12种关系对，分隔后的结果如图3所示。

基于药物与不良反应的二元组，本文使用了关联规则算法，计算二元组的置信度构建数据集进行下一步实验计算，并过滤掉置信度较小的二元组，生成的数据集合中，药物与不良反应二元组的数量与置信度大小成正比，该数据构建方式大大扩充了不良反应挖掘的实体对象范围，解决了普通挖掘方式只针对某几种药物或不良反应的局限性。

图 3分隔后的药物与不良反应二元组



1. **分布式向量的生成与关联度计算**

Google公司发布的开源工具word2vec[4-6]已经得到广泛的使用，其中Skip-gram与CBOW模型可以利用滑动窗口捕捉词语的共现信息，生成相应的高纬度分布式向量。相比于普通的one-hot方式生成的向量，该方法能够学习词语之间的语义与语法信息，更好的反映词语之间的关系，因而被广泛的使用在聚类，分类，同义词发现，关系抽取等多种自然语言处理任务中，本文将该模型应用在药物潜在不良反应挖掘领域，能够有效的识别潜在的未知关系，扩大识别范围。

本文将上述的药物与不良反应二元组作为一个实体集，进行基于分布式向量训练，其中实体集中的两个实体不具备先后顺序。训练开始前，每个药物与不良反应随机生成对应的向量，然后根据两个实体的共现关系，调整两个实体的向量内积，最终生成对应的分布式向量。实验中发现，若某不良反应只在某一种或某几种药物中出现，则该不良反应对应的向量被调整的次数会很少，会影响最终不良反应发现结果。因此实验对数据中该类不良反应进行过滤，从而提高了结果的准确度，具体公式如下。

药物与不良反应二元组的集合为R，第m个二元组集合为rm，该二元组包含的药物与不良反应集合为Vm,则分布式向量训练的目标函数为：

（1）

其中条件概率为：

= （2）

最后计算关联度时候，药物ai与不良反应bj的关联度定义为两个向量的余弦相似度：

similarity(ai, bj) = cos(ai, bj) = （3）

* 1. 进度计划

**2014.12.01---2015.8.31**

1. 了解药物不良反应事件相关背景知识。
2. 学习机器学习相关理论基础与算法。

**2015.09.01---2015.12.31**

（1）阅读相关论文文献，了解业内主流方法与技术。

（2）掌握word2vec工具的使用。

（3）调研生物相关数据集。

**2016.01.01---2016.06.30**

1. 掌握数据集特征并爬取数据。
2. 进行数据清洗与标准化。
3. 进行药名与不良反应实体识别。

**2016.07.01---2016.12.31**

1. 利用关联规则算法对数据进行过滤。
2. 训练分布式词向量，并根据实验结果调整相关参数与模型。

**2017.01.01---2017.04.30**

1. 整理实验结论，完成毕业论文的撰写、修改、提交等工作。

# 2 研究内容概述

* 1. 研究进展与阶段性成果

经过前期的研究与学习，现阶段已经完成了对药物不良反应背景知识的了解，掌握了业内主流的方法与技术，并调研生物相关数据集，利用OpenFDA提供的API接口爬取3078954份不良反应报告作为原始数据，文件大小约为41.6G，过滤掉缺失值数据之后，提取有效不良反应报告数为2932598份。

在该数据的基础上，现阶段还利用自然语言处理的相关技术对数据进行了预处理与清洗，并利用MetaMap等医学文本处理工具对药名与不良反应等实现了标准化。此外，现阶段还依据SIDER数据库对数据进行过滤，提取有效实体，为实验的进行做好了数据准备。

针对数据结构的复杂与不确定性，为了提高实验精度，现阶段还正在尝试使用关联规则算法挖掘频繁项集，生成具有更高可信度的药物-不良反应二元对，并使用Word2Vec模型对实体进行分布式向量生成。

2.2创新点论述

在药物不良反应发现领域存在着以下几点不足：（1）实验大多仍使用基于统计的方法或关联规则，只能对已存在的药物与不良反应关系进行分析处理，对于潜在的未知关系适用性不高。（2）实验数据主要局限于特定种类的药物与药物不良反应，所挖掘的药物与不良反应之间的关系相对单一。（3）实验数据直接从FDA原始数据中提取，可靠性较差，且未使用已存在的医学数据库进行数据补充与完善，对实验结果产生了较大的误差影响。

本文借助医学领域相关工具对不良反应事件报告进行数据提取清洗，并利用关联规则算法过滤数据，计算相应的权重值构建药物与不良反应二元对集合，打破了现有方法大多只针对某几种特定药物的识别限制，最后使用Word2vec训练词向量，通过计算实体之间的关联度，尝试发现药物的部分潜在不良反应，并在未知潜在关系的挖掘上寻求突破，为解决该问题提供了新的思路。

# 后续工作展望

# 3.1后续工作的设想

药物不良反应发现领域已经逐步受到国内外研究领域的普遍关注，各种方法与技术也在快速发展，因此后续的工作主要从以下两个方面进行拓展研究。

**（1）数据集的扩充**

现阶段的研究方法与技术条件下，数据集的的扩充与数据集质量的提高对于实验结果的提升有着积极的影响，因此后续工作会尝试扩充现有数据集，不仅继续对OpenFDA数据集合进行扩充，也尝试使用国外论坛、用户评论等数据进行爬取，提高数据集的丰富性。

**（2）实验方法的优化**

现阶段对于数据的清洗与实体识别仍然存在一些错误，需要后续尝试使用更多的工具与方法，提升数据识别与清洗的成功率。此外，实体的分布式向量仍需要不断的调参与优化，不断的趋近数据集的分布特征，更好的反映出药物与不良反应之间的潜在关系。

# 3.2可能遇到的困难与问题

医学领域的数据集合扩充需要大量的调研与查阅过程，由于其数据集的专业性，该过程的难度也会大大提升。此外，大量数据集本身内容与结构的复杂性也带来了解析、清洗、识别的复杂度，这也是后续实验中可能遇到的问题。

此外，关联规则算法与Word2Vec模型的使用等需要更多的理论知识的学习，且模型的调参大多是超参数，需要不断的实验经验积累，根据实际效果来调优，这也对于后续工作的进行增加了难度。

# 3.3条件保障措施

学校图书馆及电子图书馆可以提供丰富的图书及文献资源，查阅便利；实验室配置有高性能计算机，能够满足数据处理方面的需求；实验室的各位师兄师姐与指导老师在整个研究过程中提供悉心的指导与研究方向的精确掌控。

# 已发表、录用的论文和已投稿的论文情况

已撰写论文《Combining Association Rules and Embedding model to Reveal Potential Adverse Drug Reactions》。

参考文献

[1] Kass-Hout T. OpenFDA: Innovative Initiative Opens Door to Wealth of FDA’s Publicly Available Data[J]. Food and Drug Administration, 2014.

[2] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular systems biology,* vol. 6, p. 343, 2010.

[3] Zeng Q T, Tse T. Exploring and developing consumer health vocabularies[J]. Journal of the American Medical Informatics Association, 2006, 13(1): 24-29.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, 2013, pp. 3111-3119.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[6] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in HLT-NAACL, 2013, pp. 746-751.

**大连理工大学学术学位硕士研究生学位论文中期报告评审意见表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学 号 | |  | | 学生姓名 | |  | 导 师 | |  | |
| 第一次中期检查 □ | | | | | | 第二次中期检查 □ | | | | |
| 导师考核意见（对学位论文内容完成情况、阶段性成果、论文进度、后续工作思路、预期目标、及论文工作存在的问题等进行考查，给出考核成绩和具体改进意见和建议）：   1. 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格 2. 是否通过：□ 通过，□ 不通过 3. 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：   导师签字：  年 月 日 | | | | | | | | | | |
| 评  审  专  家  组 |  | 姓名 | 职称 | | 学科专业 | | | 是否博导 | | 签字 |
| 组长 |  |  | |  | | |  | |  |
| 成员 |  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
| 专家组评审意见（对学位论文内容完成情况、阶段性成果是否正确、研究方案和进度是否合理、开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等进行考查，给出考核成绩，投票表决是否通过，并给出具体改进意见和建议）：   1. 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格 2. 是否通过：□ 通过，□ 不通过 3. 具体意见（可以加页）：   组长签字：  年 月 日 | | | | | | | | | | |
| 点长意见：  点长签字：  年 月 日 | | | | | | | | | | |