项目编号：_____

吉林大学“大学生创新创业训练计划”

创新训练项目

申请书

项目名称 基于Bert的医学研究文献中的细胞间相互作用研究

项目负责人 方夷回

所在学院、年级、专业 计算机科学与技术学院2018级计算机科学与技术专业

联系电话 18255938717

电子邮箱 3350925017@qq.com

指导教师姓名 张 禹 职称 副教授

填表日期 2020年06月18日

吉林大学教务处制表

填表须知

一、**本表适用于创新训练项目**。本科生个人或团队，在校内导师指导下，自主完成创新性实验方法的设计、设备和材料的准备、实验的实施、数据处理与分析、总结报告撰写等工作。

二、申报书请按顺序逐项填写，实事求是，表达明确严谨。空缺项要填“无”。

三、申请参加大学生创新训练项目团队的人数为 3—5 人。

四、申请项目，必须聘请教师作为指导老师，并请指导教师在申请书中的指导教师意见栏中签署意见。

五、填写时可以改变字体大小等，但要确保表格的样式不变；不得随意涂改；A4 纸正反面打印，左侧装订。

六、本表由项目负责人报所在学院初审，学院签署初审意见后报送教务处实践教学科（一式 3 份原件）。

七、“项目编号”由教务处填写。

八、申报过程有不明事宜，请与教务处实践教学科联系，电话 85166413。

项目名称		基于 Bert 的医学研究文献中的细胞间相互作用研究						
项目起止时间		2020 年 07 月 至 2021 年 07 月						
负责人	姓名	学院	专业	教学号	联系电话	E-mail	QQ	各类实验班
	方夷回	计算机科学与技术学院	计算机科学与技术	20181522	18255938717	3350925017@qq.com	3350925017	是☐ 否☑
项目组成员	郭海林	计算机科学与技术学院	计算机科学与技术	21181507	13233801931	1975871624@qq.com	1975871624	是☐ 否☑
	左鹏炜	计算机科学与技术学院	计算机科学与技术	20181823	17808078709	1304401801@qq.com	1304401801	是☐ 否☑
	刘一鸣	计算机科学与技术学院	计算机科学与技术	20182322	15589071676	1092087440@qq.com	1092087440	是☐ 否☑
	徐　晋	计算机科学与技术学院	计算机科学与技术	21181322	13061095873	810366146@qq.com	810366146	是☐ 否☑

指导教师				
	姓名	张禹	职务/职称	副教授
	所在单位	计算机科学与技术学院		
	联系电话	0431-85168752	E-mail	zy26@jlu.edu.cn
	对本课题相关领域研究情况	指导教师长期从事机器学习，模式识别，生物信息学方面的研究，擅长程序设计与软件架构设计。负责本课题相关自然科学基金面上项目一项。		

项目性质	1. 小发明、小创作、小设计（　）
	2. 开放实验室或实习基地中的创新性实验或新实验开发（　）
	3. 基础性研究（✓）
	4. 应用性研究（　）
	5. 社会调研　（　）
项目选题来源	1. 自主立题（✓）
	2. 教师科研课题的子项目（　）
项目学科类别	计算机科学与技术
项目受其他渠道资助情况（填“无”或具体资助来源和经费，包括获奖情况）	无

一、立项背景和依据（包括研究目的、国内外研究现状分析与评价、研究意义，应附主要参考文献及出处）

本项目的研究目的是通过使用文本挖掘和自然语言处理技术，实现一个在医学研究文献中，对细胞间相互作用相关信息的挖掘，广泛提取可能的细胞-细胞，基因-细胞间的相互作用关系（如 Hedgehog 和 Bmp 基因的共表达[1]，果蝇视觉系统内部的细胞-细胞相互作用[2, 3]、小细胞肺癌的癌细胞和癌症相关巨噬细胞的相互作用[4]），促进对复杂组织样本中细胞间相互作用的挖掘。

国内外研究现状分析与评价。人体组织中包括有上皮细胞，基质细胞，免疫细胞等多种细胞类型。不同细胞间相互作用极大的决定了复杂组织和疾病的性质[5]。细胞间相互作用在医学研究，临床应用，以及基础生物学研究中有很大意义[6]，例如在最近几年十分热门的癌症的免疫疗法研究方面[7]。目前，免疫疗法在多种癌症上取得了很好的效果，成为了转化医学研究中的重要方向[8, 9]。 免疫疗法的基本原理是通过激活癌症组织中对癌症有杀伤作用的免疫细胞（anti-tumor immunity）的机制，达到对癌症细胞的杀伤[10]。但是，在实体肿瘤的临床治疗中，很多病人对免疫疗法并没有很好的临床反应，具体原因包括癌症组织中细胞的数目过少，癌症和周围组织对免疫细胞的抑制作用，免疫细胞对癌症细胞的低识别率，以及癌症细胞自身的对免疫共计的耐受[11]。现今对免疫疗法的科研主要目标于，发现癌症细胞及其他细胞中具有的性质，来影响免疫疗法的作用[8]。

以免疫疗法为例，近期发表的的研究论文中，有很多对于实验验证的描述。例如，PD-1 和 CTLA-4 检查点的阻断在一部分患有多种肿瘤的患者中是一种有效且持久的癌症免疫疗法[12]，在 X 癌症中，Y 基因在细胞 Z 中的表达影响了，O 细胞或者疗法[13]。这其中 X、Y、Z、O 可以是不同实验中发现的结果。此类描述并不局限于免疫疗法方面的科研，可以延伸到更广泛的对细胞间相互做用的描述。这种描述提供了在特定组织中细胞间相互作用方面丰富的信息。

自然语言处理是语言学与计算机科学、人工智能等学科的交叉领域[14]，这个领域主要研究的问题是如何使用计算机程序来分析大量的自然语言[15]。命名实体识别是自然语言处理中的一项重要任务[15]。其主要目标是将命名实体从自然语言中提取出来[16]。

本课题的目标在于，开发一种对医学研究文献的文本挖掘及自然语言处理方式，达到（1）对存在描述细胞间相互作用的文献的发现及标识，（2）通过自然语言处理方法提取标识文献中细胞间相互作用的结果，以及（3）对提取的细胞间相互作用机制的存储及建

模。值得注意的是，本课题设计的医学研究文献主要包含 PubMed 数据库中文献的摘要，以及 PubMed Central 中开源的全文本。我们将主要面向人类及小鼠系统，但目标文献并不局限于特定的组织，疾病或研究方向。同时，我们也将探索国内不同自媒体对科研结果的报道，来辅助提取科研文献中细胞间相互作用的信息。本课题主要的结果包括（1）从科研文献中提取细胞间相互作用信息的方法和（2）从文献中提取的细胞间相互作用。

本项目的**研究意义**在于认识生物本质，了解细胞间相互作用的可能机制及语言描述，阐明生物信息之间的关系；改变生物学的研究方式，引进了现代信息学和自然语言处理方法；免除大量的人类计算及知识提取工作，根据大量的数据能够直接快速地总结细胞间的相互作用。

主要参考文献

1. Bitgood, M.J. and A.P. McMahon, *Hedgehog and Bmp genes are coexpressed at many diverse sites of cell-cell interaction in the mouse embryo.* DEVELOPMENTAL BIOLOGY-ACADEMIC PRESS-, 1995. **172**: p. 126-126.
2. Reinke, R. and S.L. Zipursky, *Cell-cell interaction in the Drosophila retina: the bride of sevenless gene is required in photoreceptor cell R8 for R7 cell development.* Cell, 1988. **55**(2): p. 321-330.
3. Banerjee, U. and S.L. ZipurskY, *The role of cell-cell interaction in the development of the Drosophila visual system.* Neuron, 1990. **4**(2): p. 177-187.
4. Iriki, T., et al., *The cell-cell interaction between tumor-associated macrophages and small cell lung cancer cells is involved in tumor progression via STAT3 activation.* Lung Cancer, 2017. **106**: p. 22-32.
5. Nusse, R., *Wnt signaling in disease and in development.* Cell research, 2005. **15**(1): p. 28-32.
6. Singer, S.J., *Intercellular communication and cell-cell adhesion.* Science, 1992. **255**(5052): p. 1671-1677.
7. Bell, B.M., et al., *Designer exosomes as next-generation cancer immunotherapy.* Nanomedicine: Nanotechnology, Biology and Medicine, 2016. **12**(1): p. 163-169.
8. Rosenberg, S.A., J.C. Yang, and N.P. Restifo, *Cancer immunotherapy: moving beyond current vaccines.* Nature medicine, 2004. **10**(9): p. 909-915.
9. Mellman, I., G. Coukos, and G. Dranoff, *Cancer immunotherapy comes of age.* Nature, 2011. **480**(7378): p. 480-489.
10. Dranoff, G., et al., *Vaccination with irradiated tumor cells engineered to secrete murine granulocyte-macrophage colony-stimulating factor stimulates potent, specific, and long-lasting anti-tumor immunity.* Proceedings of the National Academy of Sciences, 1993. **90**(8): p. 3539-3543.
11. Whiteside, T.L., *The role of immune cells in the tumor microenvironment*, in *The Link Between Inflammation and Cancer*. 2006, Springer. p. 103-124.
12. Baumeister, S.H., et al., *Coinhibitory pathways in immunotherapy for cancer.* Annual review of immunology, 2016. **34**: p. 539-573.
13. Viswanathan, V.S., et al., *Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway.* Nature, 2017. **547**(7664): p. 453-457.
14. Cambria, E. and B. White, *Jumping NLP curves: A review of natural language processing research.* IEEE Computational intelligence magazine, 2014. **9**(2): p. 48-57.
15. Manning, C.D., et al. *The Stanford CoreNLP natural language processing toolkit*. in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.
16. Mohit, B., *Named entity recognition*, in *Natural language processing of semitic languages*. 2014, Springer. p. 221-245.
17. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.
18. Mikolov, T., et al. *Recurrent neural network based language model*. in *Eleventh annual conference of the international speech communication association*. 2010.
19. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

二、项目研究内容（项目主要研究内容；拟解决的关键问题、重点和难点）

本项目的**主要研究内容**包括：1. 收集相关医学研究文献；2. 抽取医学研究文献中的细胞、基因等实体；3. 提取医学研究文献中实体的相互作用关系；4. 挖掘复杂组织样本中细胞间的相互作用。

对医学研究文献中实体的相互作用关系的提取以及复杂组织样本中细胞间的相互作用是我们**拟解决的关键问题**。我们研究的**重点**是给出一种通过文献挖掘得到复杂组织样本中细胞间相互作用的方法和数据，而**难点**主要体现在如何准确的提取医学研究文献中实体的相互作用关系上。

三、项目特色及创新点

为了了解复杂组织样本中细胞间的相互作用，传统的方式是阅读大量的文献，而人类阅读文献耗时甚巨。本项目计划使用文本挖掘和自然语言处理技术来解决人类无法同时阅读海量文献的问题，这是本项目的**特色与创新点**。

四、申请理由（1、团队条件——自身/团队具备的知识、素质、能力、特长、兴趣；2、前期准备基础等）

团队条件。项目团队均是计算机科学与技术学院的学生，团队成员学习勤奋刻苦、认真努力，GPA 均不低于 3.5，同时对于计算机语言的掌握优于其他学院，有着先天的编程优势。部分团队成员曾获国家奖学金、国家励志奖学金。负责人和部分团队成员曾参加过类似项目的设计，具有相关经验，可以明确分工协作来解决实际问题。部分团队成员算法设计上独树一帜，有望在项目运行过程中设计出新的算法。另外，团队成员有着将计算机科学技术与实际的生物学结合的强烈兴趣。团队氛围佳，成员心态好，无论是遇到困难还是别的都会相互交流，共同面对，“艰难困苦，玉汝于成”是我们的信仰，“乘风破浪直挂云帆”是我们的态度。

前期准备基础。团队成员已经学习并熟悉了生物信息学相关概念，算法和相关术语。前期学习并掌握了 Word2vec 模型[17]和循环神经网络[18]，了解了词向量的建模方式。同时，对 Bert[19]中的源码有一定程度的了解。团队成员学习了一些数据库的相关知识，观看了数据挖掘的在线课程。

五、项目实施方案（研究思路和方法，实施计划、技术路线、人员分工等）

很多医学研究文献中包含了各种细胞-细胞相互作用、细胞-基因相互作用的描述，但是阅读这些文献，并从中提取出的细胞间相互作用机制需要耗费大量的时间。而最新的自然语言处理技术可以相当准确的进行命名实体识别。将最新的自然语言处理技术和对细胞间相互作用的文献提取结合起来，可以得到我们的**研究思路和方法**：使用 Bert 等自然语言处理技术对 PubMed 等医学研究文献中的细胞、基因等命名实体进行识别，分析其相互作用关系进而挖掘出细胞间相互作用的机制。

实施计划和**技术路线**如下：

1. 通过对医学研究文献的整理，生成一定范围的训练集；

2. 基于 Bert[19]，完成命名实体识别对细胞间相互作用描述语义的训练；

3. 在大量文献中测试方法，并总结所提取的细胞间相互作用机制，及人工的结果检验。

人员分工如下：

左鹏炜负责领导项目组成员进行医学相关文献收集；

刘一鸣负责领导项目组成员进行文献整理与训练集生成；

徐晋负责领导项目组成员使用 Bert 完成命名实体识别及对细胞间相互作用描述语义的训练；

郭海林负责领导项目组成员对所提取的细胞间相互作用机制进行总结，同时负责领导项目组成员进行人工的结果检验；

方夷回负责领导项目组成员进行大量文献测试与报告编写。

六、项目进度安排（文献查阅、社会调查、方案设计、开题报告、实验研究、数据处理与分析、研制开发、填写结题表、撰写论文和研究报告、结题答辩和成果推广等时间安排）

2020 年 4 月~7 月：文献查阅。了解细胞-细胞相互作用、细胞-基因相互作用；熟悉 Bert 等自然语言处理技术。5 月~6 月：方案设计。确定项目进度安排与预期成果。6 月：开题报告。确定项目研究内容与实施方案。7 月~8 月：实验研究。尝试在小规模或者模拟医学研究文献数据上进行实体抽取。8 月~9 月：数据处理与分析。挖掘细胞-细胞相互作用的信息。8 月~10 月：撰写论文。给出一种挖掘复杂组织样本细胞间相互作用的流程（pipeline）。8 月~12 月：登记软件著作权。10 月~2021 年 4 月：研制开发。

2021 年 5 月：撰写研究报告。6 月：填写结题表，结题答辩。7 月：成果推广。

七、项目研究所需资源（实验室、仪器设备、实验材料、资料等）

1. 高性能（GPU）计算平台；
2. 配置集成开发环境的工作站；
3. 国际互联网接入；
4. 包含细胞-细胞相互作用的医学研究文献与其他项目相关的各种资料。

八、项目经费预算与用途（购置实验消耗材料、低值品、资料、加工测试、打字复印、调研、市内公交、论文发表、专利申请等经费开支）

1. 资料费 1000 元，用于书籍等相关资料的购买、检索与租借等费用；
2. 打字复印费 500 元，用于项目中的文档工作；
3. 论文发表费 5500 元，用于论文发表的版面及其他相关费用；
4. 其他费用 3000 元，用于软件著作权申请、服务器租用或购买以及上述未能涵盖的费用。

九、项目完成预期成果（成果形式：研究论文、专利、设计、产品、软件、研究或调研报告等）

1. 撰写项目相关的研究论文一至二篇，并投稿至 SCI 刊物；
2. 完成数据库设计一项，并在互联网上公开发布，供相关研究人员使用；
3. 开发项目相关的软件二至三款，并登记软件著作权。

十、项目诚信承诺

本项目负责人和全体成员郑重承诺：该项目研究不抄袭他人成果，不弄虚作假，按项目研究进度保质保量完成各项研究任务。

项目负责人签名：

2020 年 06 月 18 日

项目组成员签名：

2020 年 06 月 18 日

十一、指导教师意见（从项目科学性、前沿性、可行性、研究性、可操作性和成效性进行评价，是否同意立项）

项目符合客观事实的标准，吸收了自然语言处理最新的研究成果，实验室现有条件满足其要求；其要解决的问题是对细胞间相互作用的文献提取结合，可以应用学生学到的知识，且问题定义明确；解决方案可靠，对项目预期成果的定义明确。

同意立项。

签 名：

2020 年 06 月 18 日

十二、学院评审意见（学术价值、预期效果、研究方案可行性、是否同意立项）

工作组组长签名（公章）：

年 月 日

十三、学校意见

年 月 日