**复旦大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 黄呈松 | | 学号 | 19302010004 |
| 所在院系 | 软件学院 | | 专业 | 软件工程 |
| 指导教师 | 郑骁庆 | | 职称 | 副教授 |
| 校外指导教师  及其所属单位 | 无 | | 职称 | 无 |
| 论文（设计）题目 | 基于语言模型的日常任务规划 | | | |
| **开题报告：建议包含以下内容（可另附页）：**  1. 选题的目的和意义；  2. 国内外相关研究状况综述（列出相应的参考文献）；  3. 主要研究内容与基本思路，详细技术路线，并分析可行性、难点和创新点；  4. 预期成果及形式。  见附页 | | | | |
| **研究进度及具体时间安排（不够写可加行）** | | | | |
| 起止日期 | | 主要研究内容 | | |
| 2023.02.07-2023.03.03 | | 阅读国内外相关文献，撰写开题报告 | | |
| 2023.03.04-2023.03.31 | | 编写实验代码，记录实验数据 | | |
| 2023.04.01-2023.04.21 | | 分析实验数据，验证实验结果 | | |
| 2023.04.22-2023.05.05 | | 撰写论文，定稿提交 | | |
| **指导教师对课题报告的意见：**  **1．对选题依据、基本思路或技术路线的可行性、创新性的评价；**  **2．存在的主要问题和改进建议。**  指导教师签名： 年 月 日 | | | | |

# 开题报告

研究背景

日常机器人是一个正在发展的领域。随着人口老龄化的加剧，更加需要可以在家中完成一系列日常任务的机器人。而将人工智能应用在这类机器人中也是十分有前景的方向。已有一些工作通过人工智能算法来操控机器人在房间内完成这些工作，这些算法主要通过摄像头看到的图片与文本指令来执行相关动作。

而在这一过程中，规划是至关重要的。以我们给机器人下令“泡一杯茶”为例，我们期望机器人自动的完成找到“茶杯茶叶，找到热水，泡好茶，端上来”这整个过程，而不是需要用户分四次布置这样的指令。这要求我们可以通过“泡一杯茶”这一全局指令来生成详细的四步规划。

预训练语言模型在不同的任务中都取得了不错的性能，例如问题回答和常识推理等。这证明了预训练语言模型拥有一些解决这些问题所需要的常识知识，例如微波炉可以加热食物。已有一些工作研究预训练语言模型是否可以将这些知识用于日常任务的规划中。但受限于预训练语言模型没有观察外部世界和与外部世界交互的能力，先前的工作并没有将环境信息纳入考虑。使得这些任务的规划中缺少例如“向左转”类似的导航规划。而这些规划是机器人可以适应新的环境的关键因素。

相关研究综述

**作者姓名消歧任务**

本节介绍了与作者姓名消歧相关的研究。在[3]中，作者将现有的作者人名消歧算法分为两大类，即作者分配（author assignment）和作者分组（author grouping）。作者分配指的是，把每篇论文直接与真实世界的作者一一对应起来，作者的姓名相当于论文的标签。然而，这种方法很难实现，因为它需要预先知道真实世界的所有文献作者的名称，但这几乎是不可能的。作者分组指的是，通过相似函数对同名作者对应的文献进行聚类，得到与真实世界对应的结果。作者分组方法不需要预先知道作者的姓名，且与真实世界的作者数量无关，因此在大多数情况下更容易实现。[16]总结了作者姓名消歧任务的范式：对文献依据作者姓名进行分块——在块内两两比较文献的相似度得到相似度矩阵——在相似度矩阵上运用聚类算法。提升消歧的准确性一般从以上三个步骤中选择一个或多个进行优化。

我将作者姓名消歧算法大致分为三大类：基于规则的消歧算法、有监督消歧算法、无监督消歧算法。

**基于规则的消歧算法**

基于规则的方法采用一组人为预定义的规则来考虑数据集中的每两篇文献是否属于同一作者。在[4]中，作者提出了一种基于规则的分类器，该分类器将两篇属于同名作者的文献作为输入，并根据这两篇文献的属性（如标题，共同作者列表，引用作品等）的重叠度为每个属性打一个相似度分数。对所有同名作者的文章两两进行如上操作，可以得出一份属性相似度列表，最终人为地划定阈值，确定消歧结果。尽管这种方式很简单，易于操作，但是这种方法的性能比较低下，很难在不同的领域中推广。

GHOST[5]是另一种建立在图模型之上的基于规则的方法。它为每一篇文献 构造了一张“共同作者图”模型，图中的每个节点是所有具有相同姓名的作者。只需要对数据库中的所有文献都进行相似的操作，就可以得到一张多层次的巨大的“共同作者图”，称为“结果图”。最后，通过计算结果图中每两篇文献之间的有效路径的长度，标识相似度，并使用Affinity Propagation[6]聚类算法，完成消歧。但是，这种方法不适用于只有1名作者的文献，并且没有考虑其他文献的属性(如标题、摘要或关键字)，丢失了很多关键的信息。

**有监督消歧算法**

有监督消歧算法系统地考虑了文献的属性，如标题、关键词、共同作者、地点、年份等。此类算法尝试从手工标注的训练数据中为每个作者学习特定的消歧规则。训练数据的格式一般为“作者——该作者对应的所有文献”。

在[7]中，作者提出了两类有监督消歧算法，分别是基于朴素贝叶斯算法（Naïve Bayes）与支持向量机（Support Vector Machines）[8]。朴素贝叶斯这样的生成式统计模型在学习过程中只需要正确的样本，不需要进行错误样本的修正。所以，这类方法只能对数据集中作者数量为一位的文献进行训练，在训练过程中，再运用其他可以获得的信息，进行消歧。与此相反，支持向量机是判别式模型，它的目标是训练一个分类器，用以分别文献是否对应某位作者。在这种情况下，支持向量机的训练需要用到身份不同的同名作者撰写的文献以及实际作者撰写的文献。

此外，[9]中提出了训练一个基于规则的分类器模型，将规则与有监督方法结合在了一起。这个方法把文献属性（如标题、摘要或关键字）上的某些关联规则用于推断发表该文献最有可能的作者。在后续的训练过程中，通过使用可靠的预测方法和检测训练数据中缺失的作者，将新的“负面”样本加入到训练数据集中，变相的做了数据增强。

进入神经网络时代，[10]率先提出了用深度学习的方式解决作者的姓名歧义问题。[10]提出了一种利用Network Embedding学习到的文献之间的关系信息，来处理“文献匿名图”的方法。“文献匿名图”指的是，该方法为文献数据集构造的三个局部图：表示作者之间合作关系的人-人图、表示作者与文献之间关联的人-文献图、基于合作者关系的文献-文献相似度图。它提出了一种新的针对文献的表示学习框架，通过两两文献的相似度排序，设计联合目标函数，对神经网络进行训练，将各幅图的节点嵌入到共享的低维空间中。最后的消歧结果通过聚类层次聚类得到。虽然这种特别适用于下游的聚类任务，但是由于这种方法是为匿名图设计的，因此它没有考虑文献的其他属性，而是通过简单的连接、共享来计算两篇文献的相似度。

另外，[11]提出了更优秀的文献表示学习方法。它基于数据库的全局信息与文献局部信息，优化了每篇文献在向量空间中的表示，具体方法是：利用全局信息为每个文档创建Embedding表示，再加上每篇文献与其他文献的链接的局部信息。在进行层次聚类之前，它会使用循环神经网络估计每个类的大小，做到充分减小误差。该模型是迄今为止最复杂的，这也使得它的表现优于以前的所有模型。然而，该方法需要大量的人工标注样本，并且需要涉及复杂的特征工程，不具有很高的迁移性。

针对文献难以用向量有效地进行表示的问题，[12]使用了在自然语言处理领域性能非常优异的预训练模型BERT[13]，修改了预训练任务，提出了专门针对文献表示的预训练模型SPECTER。由于参与预训练的数据量非常巨大，SPECTER甚至不需要在下游任务的数据集上进行微调，就能很好地表示文献。基于此，SPECTER在文档级别的任务上，也取得了非常良好的效果。

此外，还有一些最新提出的新型有监督算法，比如[14][15]提出了基于知识图谱（Knowledge Graph）构建文献的表示向量，在进行层次聚类以消歧聚类的算法。这种方法需要使用大规模的知识图谱数据，方便将模型迁移到新增的论文中。但是目前还没有公开的大规模文献知识图谱数据集。

**无监督消歧算法**

有监督消歧算法尽管性能良好，但是缺点却非常明显：高昂的计算成本、耗时巨大的训练过程、无法迁移的数据集等。这些不足之处使得一部分学者将无监督的方法运用到这个任务上。

[17]提出了以两篇文献的引用文献中共同作者出现的频率，作为文献相似性的方法。这个方法虽然非常简便，但是有非常多的文献信息未曾使用，导致了信息的浪费。[18]提出了在Web服务上搜索文献标题，将搜索引擎返回的网站地址的集合视为向量的方法。[18]借鉴自然语言处理中逆文档频率（Inverse Document Frequency）的概念，创新性地提出了“逆客户端频率”（Inverse Host Frequency）的概念，即在检索得到的网站地址集合中，赋予相关程度较高但却非常稀有的网页以较高的权重，例如作者本人的网页。进而，根据从“逆客户端频率”较高的网页中获得的附加信息，以层次聚类（HCA, Hierarchical Clustering Algorithm）的方式进行消歧。

目前的无监督消歧算法还没有取得很好的性能，但是其低计算成本、可增量式消歧的优点是不容忽视的。

**发明人姓名消歧任务**

本节介绍了针对专利的发明人姓名消歧算法。这一领域的工作相对于作者姓名消歧任务较少，甚至没有可用的数据集。

[19]指出，从专利数据中提取有用信息的一个主要障碍是发明人姓名的歧义问题，这导致专利系统无法将个人或机构与实体对应起来，也就无法唯一地确定参与知识生产和传播的利益相关方。它提出了一种新的算法，该算法使用高分辨率的地理位置来消除大约850万项专利的发明者和受让人的歧义，这些发明专利来自于欧洲专利局（EPO），专利合作条约（PCT），美国专利和商标局（USPTO）。虽然它的结果非常优秀，但是高分辨率的地理位置、来自官方的专利集合都是一般学者无法单独拿到的，除非有来自政府的支持。

[20]对自然语言处理在专利这一知识产权领域的应用做了一个回顾。此外，它在专利领域引入了深度学习，通过数据流、算法、用户界面、描述性统计和一个特定词在专利语料库中首次出现的位置，构建新式度量标准，自动消除专利的发明人歧义。

**参考文献**

1. Tang J, Zhang J, et al. ArnetMiner: extraction and mining of academic social networks. // Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). 2008: 990-998.
2. 中国国家知识产权局，欧洲专利局，日本特许厅，韩国特许厅，美国专利商标局. 2019年世界五大知识产权局统计报告（中文版）// url: https://www.cnipa.gov.cn/module/download/down.jsp?i\_ID=172311&colID=90. 2020.
3. Ferreira, Anderson A. et al. A brief survey of automatic methods for author name disambiguation. [J] // SIGMOD Rec. 41. 2012: 15-26.
4. Emiel Caron, Nees-Jan van Eck. Large scale author name disambiguation using rule-based scoring and clustering: International conference on science and technology indicators. [C] // Proceedings of the Science and Technology Indicators Conference 2014. 2014:79-86.
5. Fan Xiaoming, Wang Jianyong, et al. On Graph-Based Name Disambiguation. [C] // ACM J. Data Inf. Qual. 2. 2011.
6. Frey, Brendan J, Delbert Dueck. Clustering by Passing Messages Between Data Points. [J] // Science 315. 2007: 972 - 976.
7. Han H. et al. Two supervised learning approaches for name disambiguation in author citations. [C] // Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. 2004: 296-305.
8. Cortes Corinna, Vladimir Naumovich Vapnik. Support-Vector Networks. [J] // Machine Learning 20. 2004: 273-297.
9. Veloso Adriano et al. Cost-effective on-demand associative author name disambiguation. [J] // Inf. Process. Manag. 48. 2012: 680-697.
10. Zhang Baichuan, Mohammad al Hasan. Name Disambiguation in Anonymized Graphs using Network Embedding. [C] // Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.
11. Zhang Y, Zhang F, et al. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop[C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.
12. Arman Cohan, Sergey Feldman, Iz Beltagy, et al. SPECTER: Document-level Representation Learning using Citation-informed Transformers[C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2270–2282
13. Kenton J D M W C, Toutanova L K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] // Proceedings of 2019 NAACL-HLT. 2019: 4171-4186.
14. Santini Cristian, Genet Asefa Gesese, et al. A Knowledge Graph Embeddings based Approach for Author Name Disambiguation using Literals. [J] // ArXiv abs/2201.09555. 2023.
15. Färber Michael, Lin Ao. The Microsoft Academic Knowledge Graph Enhanced: Author Name Disambiguation, Publication Classification, and Embeddings. [J] // Quantitative Science Studies. 2023.
16. Shivashankar Subramanian, Daniel King, Doug Downey, et al. S2AND: A Benchmark and Evaluation System for Author Name Disambiguation[C] // 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2021: 170-190.
17. Tan Y, Kan M, and Lee D. Search engine driven author disambiguation. [J] // Proc. of 2006 ACM/IEEE Joint Conference on Digital Libraries. 2006: 314-315.
18. Kang I., Na S.H., et al. On co-authorship for author disambiguation. [J] // Journal of Information Process and Management, vol. 45, no. 1. 2009: 84–97.
19. Morrison Greg, et al. Disambiguation of patent inventors and assignees using high-resolution geolocation data. [J] // Scientific Data 4. 2017.
20. Balsmeier Benjamin, et al. Machine Learning and Natural Language Processing on the Patent Corpus: Data, Tools, and New Measures. [J] // ERN: Other Micro-economics: Intertemporal Firm Choice & Growth. 2018.

**研究内容及意义**

**研究内容**

本文研究并提出一种新型的针对文献与专利的姓名消歧算法，包含三大模块：1）对文献的作者姓名消歧；2）对专利的发明人姓名消歧；3）将文献消歧结果与专利消歧结果进行合并，可以获得某位学者名下的所有文献与专利。最主要的研究内容集中在对文献消歧结果与专利消歧结果的合并上。

**研究意义**

在大数据时代，人力资源管理越来越依靠对数据的运用，如何运用已有数据科学地评估一名人才成为了当下火热的话题。本文认为，对学术型人才的评估可以从文献和专利两个方面下手，因为二者都是可以查询获得的研究成果。因此，为了更加精确地进行人才评估，将文献和专利进行消歧、合并是有很必要的，只有这样，才能将对应的科研成果精确地匹配到相应的学者身上。

**可行性、难点和创新点**

可行性：尽管针对文献的作者姓名消歧任务与针对专利的发明人姓名消歧任务已经获得了越来越多的关注，也有了比较多的研究成果，但是这两个领域在外界看来依然是相对独立的。但是我认为，文献数据与专利数据的表示形式都是非常结构化的，方便大规模存储

难点：本文是该领域第一次将规则、无监督与有监督三种范式结合来解决作者姓名消歧任务的文章，没有相应的Benchmark。但是可以通过消融实验，显示各个部分对性能提升的贡献。此外，本文也是第一次将文献消歧结果与专利消歧结果进行合并，获得某个人的成果的工作，目标在于科研成果的跨域整合。

创新点：对规则的精心设计，使得文献在被聚类之前可以得到充分的预处理，最大限度地保证了聚类的正确性，减少不必要的错误。与此同时，本文引入了基于BERT的，针对于文献数据表示的预训练模型——SPECTER编码器，使得对文献的表示更加到位。最后，本文运用了层次聚类的无监督算法，使得聚类效果有了很大的提升。

**预期成果**

预期成果为一种新型的针对文献与专利的姓名消歧算法，包含三大模块：1）对文献的作者姓名消歧；2）对专利的发明人姓名消歧；3）将文献消歧结果与专利消歧结果进行合并，可以获得某个人的所有研究成果（文献与专利）