单位代码 10006

学 号 18231098

分类号

密 级 公开

****

毕业设计(开题报告)

**—— 基于深度学习的专利与标准的关联关系发现技术**

|  |  |
| --- | --- |
| 院（系）名称 | 高等理工学院 |
| 专业名称 | 计算机专业 |
| 学生姓名 | 霍飞烨 |
| 学院指导教师 | 张辉 |
| 工业部门导师 |  |

2020年2月25日

# 1.论文选题依据

## 1.1研究背景

在前一段时间，世界知识产权组织（WIPO）公布了 2021 年PCT国际专利申请排名榜单，申请量排名前 5 位的国家分别是中国、美国、日本、韩国和德国，其中中国企业专利数量共计 69540 件，同比增长 0.9%，连续第三年位居申请量排行榜第一位。爆炸式增长的专利背后，是繁复严谨的技术标准的支持与约束。但是随着新的技术领域的不断开拓，旧有的技术标准面临着更新与完善的问题。这就需要我们分析与相关领域的专利相关联的技术标准。

在此背景下，山西省“十四五”规划和山西省委《关于实施“111”创新工程支撑引领高质量转型发展的意见》（晋办发[2020]5号）、山西省政府《实施“111”创新工程支撑引领高质量转型发展工作方案》（晋政办发[2020]27号）文件精神，聚焦“111”创新工程的标准化需求，全面调研国内外标准化及专利现状，研究形成科技创新所需的标准化及专利数据集成方案，集成与产业密切相关的标准化及专利数据信息，产出1-2个优势重点技术领域的标准、专利内容揭示数据集，提出“111”创新工程的标准及专利创制服务体系建设方案，通过构建知识重组能力来加快技术标准及专利形成，以标准数据及专利数据形成的知识库服务促进山西省科技成果产业化，以标准化服务助力山西省打造一流创新生态，为实现山西省经济社会高质量转型发展提供标准化支撑。

传统的标准化分析方法，需要相关的专家人工分析与新增专利的相关联的技术标准，从而再根据提炼的结果来制定相关的决策。在过去专利数量增速未曾如此迅猛的情况下，这种方法未尝不是一种解决问题的有效方式。但是随着信息时代的到来，与专利相关的信息也正式步入大数据时代，此时仍然依赖人工分析数据并制定决策便显得低效且不可避免地导致决策准确率的下降。因此，我们亟需一个高效且准确的方法来对日新月异的技术专利与现有的技术标准进行数据挖掘。

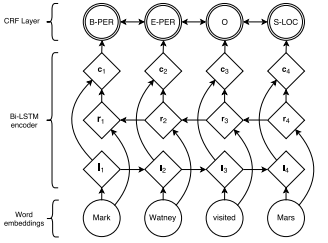
## 1.2研究意义

在深度学习，尤其是其中NLP领域的算法尚未完全成熟的过去，分析某个技术领域的专利与标准的关联，需要人工整理数据并提取其中的关系。由于效率限制，支撑结论的数据量通常不会太过庞大，且耗时巨大。而通过爬虫与深度学习相结合，则可以在短时间内大量提取特定领域的相关专利中关联的技术标准，从而更加高效准确地做出决策判断。

本毕设以服务“111”创新工程标准化需求为导向，跟踪国内外先进技术专利信息，负责对专利数据与标准数据进行关联研究，指导企业及时跟踪科技创新前沿，寻找专利申请的主要方向，开展精准创新。基于科技创新前沿专利画像模型、标签体系及标准数据，抽取专利实体的特征标签，挖掘专利、标准实体关系。

## 1.3国内外研究现状

国内外对于专利与标准的关联关系的发现这一具体的子课题的研究并不多。但是对专利与标准的关联关系的发现问题，本质上就是自然语言处理（NLP）中的命名实体识别（NER）课题。而NER是一个具有挑战性的学术课题。一方面，在大多数语言和领域，只有非常少的监督训练数据可用。另一方面，对于可以作为名称的单词种类几乎没有什么限制，因此从这个小数据样本中进行概括是困难的。因此，许多性能优秀的模型都严重依赖于人工构造的特征。为了解决这个问题，Carnegie Mellon University的Guillaume Lample、Miguel Ballesteros、Sandeep Subramanian、Kazuya Kawakami以及Chris Dyer在2016年发表的论文《Neural Architectures for Named Entity Recognition》中提出LSTM+CRF模型。它不使用任何特定于语言的资源或特性，只使用少量有监督的训练数据和未标记的语料库，是一种端到端的判别式模型。LSTM 利用过去的输入特征，CRF 利用句子级的标注信息，源序列先输入到LSTM中，然后输出一个单词的考虑到上下文的特征到CRF中，因为CRF可以考虑全局依赖。因此，该模型可以有效地使用过去和未来的标注来预测当前的标注。以序列Mark Watney visited Mars为例，其大体的模型如下：



# 2.研究方案

## 2.1研究目标及研究内容

2.1.1 研究目标

本毕设，旨在开发一个后端包括以技术专利与相关技术标准关联关系为数据挖掘对象的深度学习算法模型，并且结合爬虫技术来爬取最新的专利信息作为其数据对象。同时，由于本毕设面对的预期用户群体，是需要根据新专利与旧标准间的关联关系进行分析的决策者，因此本毕设还需要设计一个面向用户的友好的可视化前端。

本毕设的成果的应用场景，是通过深度学习方法，分析大批量的专利信息，抽取其中的相关技术标准，并且通过可视化的前端表现，从而辅助决策者是否要在某个技术领域补充新的技术标准。

2.1.2 研究内容

在研究内容方面，本次毕设将会涉及深度学习中NLP方面的技术、爬虫技术以及前后端开发技术，因此，本毕设并非是聚焦于对现有的深度学习算法进行性能改进的学术性研究，而是希望将已经成熟的相关技术应用于专利与标准间关联关系的发现的工程开发。

本毕业设计的主要难点将会是如何将成熟模型框架运用于具体的问题——即专利与标准间关联关系的发现，因为本毕设的主要目标是为了开发效率更高、精度更准的模型，在这过程中，不仅仅需要选取合适的算法模型，对于某一个具体的算法模型，还需要不断地修改、调整参数以及训练，这个过程将会耗费大量的数据以及时间；其次，为了使得分析结果具有实时性，我们还需要爬虫技术来爬取最新的专利技术的信息，因此，如何快速爬取正确的数据也将会是影响成品性能的重要方面，与此同时，爬虫的性能也会受到许多来自非学术性原因的网络安全问题的约束，因此，我们不仅仅要解决爬虫过程中的性能问题，还需要解决其成功与否的问题；最后，作为一个工程产品，毕设成果也必须具有一定的可视性与用户交互性，当然，由于本毕设的主要矛盾在于分析结果而非用户的使用体验，因此，关于这一方面的内容不会过于深入，只会开发一个简易的可视化前端，但是最终成品是否要进行图表化的显示、是否要进行对于轻量级软件来说相对复杂的前后端分离式开发也是一个等待具体开发过程中进行抉择的一个难点问题。

## 2.2 拟采用的研究方法、技术路线

2.2.1 研究方法

本次毕设，将会采用文献研究法、比较研究法与实证研究法。

文献研究法：通过图书馆、互联网、电子资源数据库等途径查阅大量文献，理解前沿NER算法模型、前沿爬虫技术等相关知识，理清NER的发展脉络及研究现状，并且围绕目前调研后选择的算法模型选择性地学习深度学习有关理论知识，为设计最终的毕设成果中深度学习模型提供思路和参照。

比较研究法：在毕设调研中，除了会确定一个主要方案外，我还会通过与老师学长的交流，敲定其余一两个备选方案，当主要方案在实现过程中发现面对这个特异性问题的性能效率并未达到预期时，将会转向备选方案进行比照实验，最终选定其中相对优越的方案作为最终方案。

实证研究法：深度学习学科是一个实验的学科，再多么优越的模型都需要大量的数据集来进行训练测试，因此，本次毕设除了一开始需要用事先准备好的静态数据集进行初试模型训练与测试调整外，在工程后期，还会结合爬虫爬取的最新数据进行实战测验，经过这两轮的实验后，方会根据实验结果来决定毕设的开发是否获得预期中的效果。

2.2.2 技术路线

本次毕设开始，我首先会与老师交流明晰整个毕设的目的任务，并且进行相应的学习。首先我会学习最基础的深度学习知识，然后由于本次毕设是NLP方向相关，因此主要的学习任务将会是循环神经网络（RNN）。在学习完最基础的深度学习知识后，我将会调研目前前沿的NLP相关课题的算法模型，并且在其中挑选出符合自己能力的同时相对优越的模型作为毕设的参照模型。随后，我会根据论文的描述或者参考其源代码搭建自己的Pytorch模型，并对其不断迭代优化，当该模型在静态数据集上表现良好后，我就会利用爬虫技术来爬取动态数据进行测试直到效果达到预期。最后，我才会开发相应的前后端并封装整个程序。

## 3. 研究计划

|  |  |
| --- | --- |
| 周 | 主要任务目标 |
| 1 | 与老师学长沟通毕设成品的主要功能，初步明确接下来的学习方向，即深度学习基础知识 --> NLP方向NER算法 --> 爬虫技术 --> 前后端开发技术。 |
| 2 | 调研前沿的NER算法论文，并听取老师学长意见，从中挑选一个在我实现能力之内的作为接下来围绕学习主要方案。同时也会挑选几个作为备用方案。 |
| 3-4 | 通过吴恩达、李沐等人工智能专家出品的广受好评的教育系列视频学习深度学习的基础知识，并且选择性地学习其中与NLP相关的部分。在掌握了相关知识后，细读理解之前调研所选中的主要方案的论文，理解算法模型的原理。 |
| 5-6 | 尝试复现、移植、匹配主要方案的模型，初步构建起模型框架，若模型存在明显的与具体问题的不适配性，则可以考虑更换备选方案。 |
| 7-10 | 敲定最终的模型框架，开始在云服务器上测试并调整、训练模型，使之适配具体的问题，最终做到在静态数据集上取得预想的性能。 |
| 11-12 | 学习爬虫技术，并且爬取一定数量的最新实时数据，由于爬虫的效果涉及许多其他非学术相关的因素，因此若该阶段的效果不理想，可能调整毕设的目标成果。 |
| 13 | 测试爬虫数据在之前训练好的模型上的效果与性能，并进行相应的调整。 |
| 14-15 | 开发相应的简易前后端，并封装整个程序，测试成品是否存在BUG。 |
| 16 | 准备最终的毕设答辩，完成毕业论文。 |

## 4. 参考文献

[1] Guillaume Lample、Miguel Ballesteros、Sandeep Subramanian、Kazuya Kawakami、Chris Dyer，Neural Architectures for Named Entity Recognition，Carnegie Mellon University ，NLP Group, Pompeu Fabra University

# [2] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. The Journal of Machine Learning Research

[3] Rie Kubota Ando and Tong Zhang. 2005a. A framework for learning predictive structures from multiple tasks and unlabeled data. The Journal of Machine Learning Research